Eine Webanwendung zur Planung und Auswertung von A/B-Tests auf Basis des Chi-Quadrat-Unabhängigkeitstests

Masterarbeit

Oliver Frost

Humboldt-Universität zu Berlin
Ladislaus von Bortkiewicz Lehrstuhl für Statistik

Betreuer
Dr. Sigbert Klinke

Gutachter
Prof. Dr. Wolfgang Härdle
Prof. Dr. Stefan Lessmann

10. März 2018

Danksagung

Ich danke meinen Eltern für ihre großzügige Unterstützung meines Studiums, ja meines Bildungsweges überhaupt. Er wäre sonst ungleich holpriger verlaufen. Ferner danke ich meinem Kollegen Marian Majewski-Sliwa. Er hat mein Interesse für A/B-Tests geweckt und mir von Anfang an die Möglichkeit gegeben, eigene Ideen in den A/B-Test-Prozess der Immobilien Scout GmbH einzubringen. Zuletzt geht mein Dank an Herrn Klinke für die gute Zusammenarbeit.

Abstract

In dieser Arbeit stelle ich eine für die Scout24 AG entwickelte Webanwendung vor, mit der sich A/B-Tests auf Basis des Chi-Quadrat-Unabhängigkeitstests planen und auswerten lassen. Die Anwendung enthält vier verschiedene Formen der Teststärkeanalyse, darunter die von Erdfelder (1984) vorgestellte Kompromissanalyse, sowie einen Auswertungsbereich, in dem neben dem p-Wert ein Effektmaß (Cohens w) samt Konfidenzintervall ausgegeben wird. Bei der Entwicklung wurde besonderes Augenmerk darauf gelegt, dass die Bedienung der Anwendung durch Personen ohne inferenzstatistisches Fachwissen erfolgen kann.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ABTP	A/B test pro
CQUT	Chi-Quadrat-Unabhängigkeitstest
DOF	Degrees of freedom
ESD	Empirische Standardabweichung
KR	Konversionsrate
MDE	Minimum detectable effect

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Konversionsrate (KR) ist eine zentrale Kennzahl im E-Commerce. Sie gibt den Anteil der Besucher einer Website an, die eine bestimmte, vom Betreiber gewünschte Handlung ausführen. Je nach betrachteter Website kann damit etwa der Kauf eines Produkts oder der Klick auf einen Werbebanner gemeint sein. Will man die KR durch eine Designänderung erhöhen, z. B. durch eine Neugestaltung des Bestellbuttons, empfiehlt es sich, vor der Umstellung zu überprüfen, ob die geplante Änderung auch wirklich den erwarteten positiven Effekt auf die KR hat.

Hierfür hat sich in den letzten Jahren ein Verfahren etabliert, das günstiger und flexibler als seine Alternativen ist: der A/B-Test. In seiner einfachsten Form sieht eine Besucherhälfte die alte Version der zu testenden Seite, während die andere Hälfte die neue Version sieht. Nach einer vorab festgelegten Zeit wertet man dann aus, welche der beiden Versionen die höhere KR hatte. Doch wie groß muss dieser Unterschied ausfallen, um mit Sicherheit sagen zu können, dass er auch auf der Designänderung beruht und nicht nur dem Zufall geschuldet ist? Die Tatsache, dass ein Unterschied beobachtet wurde, bedeutet nämlich noch nicht, dass eine Version wirklich besser als die andere ist. Selbst wenn beide Gruppen die exakt gleiche Seite gezeigt bekommen hätten, würden sie kaum die gleiche KR aufweisen.[1]

Gemäß frequentistischer Inferenzstatistik löst man dieses Problem mit einem Hypothesentest, z. B. dem Chi-Quadrat-Unabhängigkeitstest (CQUT). Solche Tests so zu planen, dass ein tatsächlich existierender Unterschied mit zufriedenstellender Wahrscheinlichkeit erkannt wird, ist allerdings kein leichtes Unterfangen und bisher ohne spezielle Software kaum zu bewerkstelligen. Zur Auswertung von Hypothesentests dagegen gibt es zahlreiche kostenlose Webanwendungen. Wie ich in Abschnitt 6.1 zeigen werde, sind diese jedoch aus mehreren Gründen problematisch.

In dieser Arbeit stelle ich eine für die Scout24 AG entwickelte Webanwendung vor, mit der sich A/B-Tests sowohl planen als auch auswerten lassen

[1] Warum dies so ist, lässt sich an einem einfachen Zufallsexperiment veranschaulichen: Wenn eine faire Münze 10 000 Mal geworfen wird, beträgt die Wahrscheinlichkeit, 5 000 Mal Kopf zu beobachten, gerade einmal 0.8%. Die Wahrscheinlichkeit für mindestens 4 950 und höchstens 5 050 Mal Kopf dagegen beträgt ganze 68.8%. Bei einer hohen Anzahl an Würfen ist es also selbst mit einer fairen Münze sehr unwahrscheinlich, eine *exakte* Aufteilung von Kopf und Zahl zu erhalten. Viel wahrscheinlicher ist es, eine geringe Abweichung zu beobachten.

und zwar ohne dass dafür tiefere Kenntnisse der Inferenzstatistik nötig wären. Die Anwendung, AB test pro genannt, basiert auf dem CQUT und ist in der freien Programmiersprache R (R Core Team 2016) geschrieben.

Zum Aufbau der Arbeit: Zunächst werde ich die zwei zentralen Begriffe dieser Arbeit – Konversionsrate und A/B-Test – näher erläutern. Anschließend folgt eine Darstellung des frequentistischen Testprinzips im Allgemeinen sowie des CQUT im Speziellen. Daraus wird deutlich, warum der p-Wert ein schlechtes Maß der Effektstärke ist und weshalb man dafür besser ein Effektmaß zu Rate ziehen sollte. Ein solches, Cohens w, wird im vierten Kapitel eingeführt und genauer betrachtet. Damit ist der Boden für das fünfte Kapitel bereitet, das sich um den Nutzen der Teststärkeanalyse und deren verschiedene Formen dreht. Zum Abschluss stelle ich im sechsten Kapitel die oben erwähnte Webanwendung vor, in der sich die Theorie der vorangegangen Kapitel widerspiegelt.

2 Zwei zentrale Begriffe

Dieses Kapitel hat den Zweck, bei Leser und Autor ein gemeinsames Verständnis der zwei zentralen Begriffe dieser Arbeit – Konversionsrate und A/B-Test – herzustellen.

2.1 Konversionsrate

Hinter jeder Website steht ein Ziel, z. B. der Verkauf von Produkten oder das Generieren von Werbeklicks. Handelt ein Besucher gemäß diesem Ziel, dann spricht man von einer Konversion. Die Konversionsrate gibt an, wie viel Prozent der Besucher in einem bestimmten Zeitraum konvertiert sind. Bei ihrer Berechnung gilt: Pro Besucher wird maximal eine Konversion gezählt (Krüger 2011: 29). Hierfür ist es notwendig, den einzelnen Besucher zu identifizieren, was üblicherweise mittels Cookies geschieht.[2]

Wie werden Konversionen technisch erfasst? Dafür gibt es im Wesentlichen zwei Methoden (Hassler 2012: 27–30):

- **Logdateianalyse:** Hier werden die Logdateien des Webservers ausgewertet. Darin ist jeder Aufruf jeder einzelnen Datei einer Website pro-

[2]Genau genommen identifizieren Cookies keine Besucher, sondern Browser. Daraus folgen Ungenauigkeiten. Greifen z. B. mehrere Personen über den selben Browser auf eine Website zu, werden sie fälschlicherweise als eine Person gezählt.

Abbildung 1: Konversionspfad aus Google Analytics

tokolliert. Die große Schwierigkeit bei der Logdateianalyse ist, einen Besucher, der mehrere Seiten einer Website aufruft, über all diese Seitenaufrufe hinweg zu identifizieren. Hierfür ist zusätzlicher Programmieraufwand nötig.

- **Page-Tagging:** Dies ist die moderne Methode der Konversionserfassung. Hier wird in jede Seite einer Website ein kleines JavaScript-Programm eingebaut, das verschiedene Informationen über den Besucher an einen externen Anbieter sendet, z. B. über welche Seite er auf die aktuelle Seite gelangt ist. Diese Daten werden dem Websitebetreiber dann online (kostenfrei) zur Verfügung gestellt.

Abbildung 1 zeigt einen von der Immobilien Scout GmbH in Google Analytics erstellten Konversionspfad. Der Pfad basiert auf Page-Tagging, erfasst fünf aufeinander folgende Seiten und zeigt, wie viel Prozent der Besucher jeweils auf die nächste Seite gegangen sind (grauer Querpfeil). Mit solchen Pfaden können Onlineshops erkennen, an welcher Stelle im Bestellvorgang Optimierungspotenzial besteht. In A/B-Tests werden sie häufig als Datenbasis eingesetzt.

2.2 A/B-Test

Wer eine kommerzielle Website betreibt, strebt danach, diese so zu gestalten, dass sie möglichst viele Konversionen generiert.[3] Dabei tauchen viele Detailfragen auf: Welche Form und Farbe soll der Bestellbutton haben? Sollen Besucher gesiezt oder geduzt werden? Wie viele Werbebanner sollen auf einer bestimmten Seite platziert werden? Anders als man meinen könnte, beeinflussen solche Feinheiten die KR mitunter erheblich. So berichten Kohavi et al. (2008: 143–44) von einem Onlineshop, dessen KR völlig unerwartet um 90% sank, als auf der Checkout-Seite ein Feld zur Eingabe eines Rabattcodes hinzugefügt wurde. Darum ist es aus ökonomischer Sicht ratsam, neue Designideen vor ihrer Implementierung stets zu testen.

Hierfür hat sich der Einsatz von A/B-Tests bewährt. Diese funktionieren wie folgt: Zunächst verändert man die Originalversion (A) einer Webseite so, dass die neue Version (B) vermeintlich das Konvertierungsverhalten der Besucher verbessert.[4] Dann stellt man beide Versionen online und zwar derart, dass neu auf die Seite kommende Besucher mit Wahrscheinlichkeit p A und mit Wahrscheinlichkeit $1 - p$ B angezeigt bekommen.[5] Typischerweise setzt man $p = 0.5$.[6] Zuletzt wertet man den A/B-Test mit einem statistischen Hypothesentest aus. Warum dies notwendig ist, warum man sich also nicht einfach für jene Version entscheiden sollte, die im Test die höhere KR hatte, werde ich in Kürze erläutern.

Der mutmaßlich erste A/B-Test wurde im Jahr 2000 von Google durchgeführt (Christian 2012). Elf Jahre später hatte Google bereits mehr als 7000 A/B-Tests durchgeführt. Heutzutage gelten A/B-Tests als *das* Standardwerkzeug der Weboptimierung. Zwei Gründe dürften für diesen Erfolg maßgeblich sein: die geringen Implementierungskosten im Vergleich zu Alternativen, z. B. Usability-Tests, und die Tatsache, dass das Testprinzip in vielerlei Hinsicht

[3] Der Einfachheit halber nehme ich in dieser Arbeit an, dass eine höhere KR gleichbedeutend mit mehr Gewinn für den Websitebetreiber ist, was meistens, aber nicht immer der Fall ist. So könnte eine bestimmte Designänderung dazu führen, dass der Anteil der Käufer zwar steigt, der durchschnittliche Kundenumsatz aber so sehr sinkt, dass die Änderung insgesamt einen negativen Effekt hat.

[4] In Abschnitt 6.4 stelle ich zwei von Immobilien Scout durchgeführte A/B-Tests vor.

[5] Die Information, welcher Version ein Besucher zugeteilt wurde, wird üblicherweise in einem Cookie gespeichert. Sofern dieser nicht gelöscht wird, sieht der Besucher daher auch bei allen nachfolgenden Besuchen die ihm ursprünglich zugeteilte Version.

[6] Wenn man sich sehr unsicher ist, wie die neue Version von den Besuchern angenommen werden wird, kann es sinnvoll sein, ein größeres p zu wählen.

erweitert werden kann. Dies sind zwei verbreitete Varianten:

- **A/B/n-Test:** Hier werden nicht zwei, sondern n Versionen einer Webseite gegeneinander getestet. Ein A/B/n-Test ist dann sinnvoll, wenn ein Gestaltungsparameter viele in Frage kommende Ausprägungen hat und die zu testende Seite stark frequentiert wird, so dass für jede Version in absehbarer Zeit eine belastbare KR ermittelt werden kann.[7]

- **Multivariater Test:** Hier werden gleichzeitig mehrere Gestaltungsparameter verändert; getestet werden dann Kombinationen verschiedener Ausprägungen. Ein Webshopbetreiber könnte z. B. testen, welche Form-Farbe-Kombination des Bestellbuttons die meisten Klicks generiert. Bei drei Farben und zwei Formen würden somit sechs verschiedene Versionen der Checkout-Seite miteinander konkurrieren. Multivariate Tests haben im Wesentlichen zwei Vorteile gegenüber einfachen A/B-Tests (Kohavi et al. 2008: 158–63): Sie beschleunigen die Optimierung der Seite, weil mehrere Änderungen gleichzeitig überprüft werden können. Außerdem können sie Abhängigkeiten zwischen den einzelnen Parametern aufdecken. Allerdings ist ihre technische Umsetzung und Auswertung ungleich komplexer, weshalb im Einzelfall genau abgewogen werden muss, ob ein multivariater Test oder eine Serie von A/B-Tests zielführender ist.

Eine Schwierigkeit haben alle Varianten mit dem klassischen A/B-Test gemein: Da den im Test beobachteten KR grundsätzlich eine Zufallskomponente innewohnt, kann allein aus dem Umstand, dass sich die beiden KR unterscheiden, noch nicht gefolgert werden, dass eine Version der anderen *wirklich* überlegen ist, denn der Unterschied könnte rein zufälliger Natur sein. Dies bedarf einer näheren Erläuterung.

Einfluss auf die KR in einer Besuchergruppe haben offenbar drei Faktoren: die Konvertierungsbereitschaft der Besucher, die angezeigte Webseite und der Zufall. Da bei einem technisch korrekt implementierten A/B-Test die Konvertierungsbereitschaft der Besucher in beiden Gruppen gleich groß sein sollte, kommen als mögliche Ursachen für einen KR-Unterschied nur die beiden letztgenannten Faktoren in Frage. Die Herausforderung bei der

[7] Ein bekanntes Beispiel für einen solchen Test stammt von Google: Dort hat man 41 verschiedene Blautöne für Werbelinks getestet, um jenen zu finden, der die meisten Klicks generiert (Hern 2014).

Auswertung von A/B-Tests besteht somit darin, zwischen KR-Unterschieden zu trennen, die allein dem Zufall, also nicht der Umgestaltung der Webseite, geschuldet sind, und solchen, bei denen sowohl der Zufall als auch die Umgestaltung das Konvertierungsverhalten der Besucher beeinflusst haben.

3 Frequentistische Auswertung

Die Auswertung von A/B-Tests ist ein Problem der Inferenzstatistik, jenem Bereich der Statistik, der sich damit befasst, wie man aus Stichprobendaten Kenntnisse über die Grundgesamtheit ableitet. In der Inferenzstatistik wiederum gibt es zwei Denkschulen: Frequentismus und Bayesianismus. Beide basieren auf grundverschiedenen Wahrscheinlichkeitsbegriffen. Frequentisten definieren die Wahrscheinlichkeit eines Ereignisses als Grenzwert der relativen Häufigkeit, mit der es in einer unendlichen Serie gleicher und voneinander unabhängiger Zufallsexperimente auftritt. Für Bayesianer dagegen entspricht die Wahrscheinlichkeit eines Ereignisses dem Grad der persönlichen Überzeugung, dass dieses Ereignis eintritt. Im Unterschied zu Frequentisten berücksichtigen sie mögliches Vorwissen des Beobachters.

Entsprechend diesen beiden Schulen gibt es bei den meisten Problemen der Inferenzstatistik zwei verschiedene Herangehensweisen, die frequentistische und die bayesianische. Die Frage, wie man einen A/B-Test auswerten sollte, bildet da keine Ausnahme. In dieser Arbeit werde ich mich jedoch ausschließlich mit dem frequentistischen Ansatz befassen.

3.1 Testprinzip

Das zentrale Instrument frequentistischer Inferenz sind Hypothesentests. Diese fungieren als Entscheidungsregel, um zwischen einer Nullhypothese (H_0) und einer konkurrierenden Alternativhypothese (H_1) zu wählen. Die Belegung der Thesen erfolgt so, dass H_1 jener Vermutung entspricht, die man zeigen möchte. Dabei handelt es sich zumeist um eine These, die über den bisherigen Wissensstand hinausgeht.

Das Eigentümliche an Hypothesentests ist, dass die Gültigkeit von H_1 indirekt gezeigt wird. Die Grundidee ist nämlich, sich erst dann für H_1 und gegen H_0 zu entscheiden, wenn die Abweichung zwischen dem, was man bei Richtigkeit von H_0 beobachten würde, und dem, was tatsächlich beobachtet wird, sehr groß ist. Man spricht dann von einem statistisch signifikanten

Ergebnis.[8]

Der Vorteil dieser indirekten Entscheidungsregel ist, dass man die Wahrscheinlichkeit α für ein fälschliches Verwerfen von H_0 (α-Fehler) kontrollieren kann. Je kleiner man allerdings diese als Signifikanzniveau bezeichnetete Wahrscheinlichkeit wählt, desto größer wird die Wahrscheinlichkeit β, zu Unrecht an H_0 festzuhalten (β-Fehler). Die Gegenwahrscheinlichkeit $1 - \beta$ wird als Teststärke bezeichnet und ist ein zentrales Gütekriterium von Hypothesentests. In Kapitel 5 werde ich erläutern, welche Faktoren die Teststärke von A/B-Tests beeinflussen und inwiefern eine zu geringe Teststärke aus Sicht des Shopbetreibers problematisch ist.

Wie lauten nun Null- und Alternativhypothese bei einem A/B-Test, in dem eine bestimmte Designänderung D überprüft werden soll? Eine naheliegende Wahl ist:

H_0: D hat keinen Einfluss auf die KR.

H_1^*: D erhöht die KR.

H_1^* wird als gerichtete Alternativhypothese bezeichnet, weil darin eine konkrete Aussage über die Richtung des Effekts von D gemacht wird. Solche Alternativhypothesen sind allerdings problematisch, wenn man wie hier mit extremen Beoabachtungen in die „falsche“ Richtung rechnen muss. Würde man nämlich in der Gruppe, die D gesehen hat, eine deutlich schlechtere KR beobachten als in der Kontrollgruppe, gäbe es keinen Anlass dafür, H_0 zu verwerfen, obwohl diese These augenscheinlich unplausibel wäre. Eine geeignetere Alternativhypothese als H_1^* ist somit:

H_1: D erhöht *oder* verringert die KR.

Der nächste Schritt ist, diese gemeinsprachlichen Thesen in Hypothesen zu überführen, die von Zufallsvariablen handeln. Werden dabei Annahmen über die Verteilungen dieser Zufallsvariablen getroffen, erhält man einen parametrischen Test; andernfalls einen nichtparametrischen Test. Bei der hier vorliegenden Frage, ob ein beobachteter KR-Unterschied statistisch signifikant ist, bieten sich aus beiden Kategorien mehrere Verfahren an. Welches also wählen?

[8] Aus der Tatsache, dass ein Ergebnis *statistisch* signifikant ist, lässt sich jedoch nicht schließen, dass es auch in praktischer Hinsicht relevant ist. Für diesen Schluss ist es notwendig, die Effektstärke zu messen. Darauf werde ich in Kapitel 4 zurückkommen.

3.2 Chi-Quadrat-Unabhängigkeitstest

Für den A/B test pro, den ich in Kapitel 6 vorstellen werde, fiel meine Wahl auf den nichtparametrischen Chi-Quadrat-Unabhängigkeitstest (CQUT) und zwar aus folgenden Gründen:

- Campell (2007) vergleicht in einer Simulationsstudie sieben verschiedene nichtparametrische Tests zur Auswertung von 2×2-Kontingenztafeln – drei Varianten des CQUT sowie vier Varianten von Fisher's exaktem Test – und gelangt schließlich zu der Empfehlung, den CQUT (genauer gesagt eine geringfügig modizifierte Version davon) immer dann anzuwenden, wenn die erwartete Häufigkeit einer jeden Zelle mindestens 1 ist. Diese Bedingung ist in der A/B-Test-Praxis üblicherweise erfüllt.

- Mit dem CQUT können nicht nur 2×2-, sondern auch $m \times r$-Tafeln ausgewertet werden, man kann ihn mithin auch zur Analyse von A/B/n-Tests und multivariaten A/B-Tests nutzen. Diese Testformen werden aktuell zwar noch nicht vom A/B test pro unterstützt, eine solche Erweiterung steht aber zur Diskussion.

- Der CQUT kommt mit nur zwei Annahmen aus, die, wie wir gleich sehen werden, bei technisch korrekt implementierten A/B-Tests in der Regel erfüllt sind.

Nun zum Test selbst. Mit dem CQUT lässt sich überprüfen, ob zwei kategoriale Merkmale X und Y mit m und r Kategorien stochastisch unabhängig voneinander sind. Als Datenbasis dient eine $m \times r$-Kontingenztafel, in der die beobachteten Häufigkeiten aller Merkmalskombinationen erfasst sind. Da der CQUT ohne Annahmen über die Verteilungen von X und Y auskommt, handelt es sich um ein nichtparametrisches Verfahren. Damit die Teststatistik approximativ Chi-Quadrat-verteilt ist mit $(m-1) \cdot (r-1)$ Freiheitsgraden, müssen lediglich die folgenden zwei Annahmen erfüllt sein:

> $A1$: Die einzelnen Beobachtungen sind voneinander unabhängig, d. h., sie beeinflussen sich nicht gegenseitig.
>
> $A2$: Die *erwartete* Häufigkeit einer jeden Zelle der Kontingenztafel ist mindestens 5.[9]

[9] Diese Forderung findet sich in vielen Statistik-Lehrbüchern, siehe z. B. Bortz & Schuster (2010: 141). Gemäß Campells oben erwähnter Studie ist diese von Cochran (1952: 328) als willkürlich bezeichnete Vorgabe allerdings überholt.

Tabelle 1: Ergebnisse eines fiktiven A/B-Tests

	A	B	
Konvertiert	24 (a)	37 (b)	61
Nicht konvertiert	103 (c)	89 (d)	192
	127	126	253

Sind $A1$ und $A2$ in der A/B-Test-Praxis erfüllt? $A1$ ist im Allgemeinen erfüllt, sofern der A/B-Test technisch korrekt implementiert ist.[10] Ob $A2$ erfüllt ist, hängt davon ab, wie stark die getestete Webseite frequentiert wird und wie lange man den A/B-Test laufen lässt. Wie ich in Abschnitt 5 zeigen werde, sollte letzteres davon abhängen, welche Laufzeit nötig ist, um eine akzeptable Teststärke zu erreichen. Diese Maxime führt dazu, dass die erwarteten Häufigkeiten in der A/B-Test-Praixs weit größer als 5 sind.

Tabelle 1 zeigt die Ergebnisse eines fiktiven A/B-Tests, wie sie beispielsweise aus einem Konversionspfad (siehe Abschnitt 2.1) stammen könnten, wobei Gruppe B der Designänderung D ausgesetzt war. Die KR in Gruppe A lag bei 18.9%, jene von Gruppe B bei 29.4%. Dies entspricht einer Verbesserung von 55.6%. Auf den ersten Blick scheint die Sache daher klar: D erhöht den Anteil der Besucher, die sich zum Kauf entschließen, und zwar deutlich. Ob die Verbesserung allerdings groß genug ausgefallen ist, um mit befriedigender Wahrscheinlichkeit ausschließen zu können, dass sie eine Zufallsschwankung ist, kann erst der CQUT klären.

Der erste Schritt bei einem Hypothesentest ist stets das Aufstellen von Null- und Alternativhypothese. Beim CQUT lauten diese wie folgt:

H_0: X und Y sind stochastisch unabhängig.

H_1: X und Y sind stochastisch abhängig.

Die Merkmale X und Y entsprechen hier der Gruppenzugehörigkeit – A oder B – und dem Besucherstatus – konvertiert oder nicht konvertiert. Um H_0 zu überprüfen, werden nun die beobachteten Häufigkeiten a–d mit jenen Häufigkeiten verglichen, die man unter Gültigkeit von H_0 beobachtet hätte.

[10] Eine Ausnahme liegt z. B. vor, wenn die getestete Designänderung sehr auffällig ist und öffentlich diskutiert wird. In diesem Fall ist nicht mehr gewährleistet, dass das Konvertierungsverhalten der Besucher unabhängig von der Gruppenzugehörigkeit ist.

Sind die Abweichungen erheblich, dann wird H_0 verworfen. Der Vergleich erfolgt mittels der folgenden Teststatistik:

$$\chi^2 = \sum_{\text{Zellen}} \frac{(\text{beob. Häufigkeit} - \text{erwart. Häufigkeit})^2}{\text{erwart. Häufigkeit}} \tag{1}$$

Die erwartete Häufigkeit einer Zelle ergibt sich, indem man das Produkt aus der jeweiligen Zeilen- und Spaltensumme durch den Stichprobenumfang teilt:

Tabelle 2: Erwartete Häufigkeiten

	A	B	
Konvertiert	$30.6 = \frac{61 \cdot 127}{253}$ (a^*)	$30.4 = \frac{61 \cdot 126}{253}$ (b^*)	61
Nicht konvertiert	$96.4 = \frac{192 \cdot 127}{253}$ (c^*)	$95.6 = \frac{192 \cdot 126}{253}$ (d^*)	192
	127	126	253

Mit a–d und a^*–d^* lässt sich jetzt ein Prüfwert für die obige Teststatistik berechnen:

$$\begin{aligned} \chi^2 &= \frac{(a-a^*)^2}{a^*} + \frac{(b-b^*)^2}{b^*} + \frac{(c-c^*)^2}{c^*} + \frac{(d-d^*)^2}{d^*} \\ &= \frac{(24-30.6)^2}{30.6} + \frac{(37-30.4)^2}{30.4} + \frac{(103-96.4)^2}{96.4} + \frac{(89-95.6)^2}{95.6} \\ &= 3.8 \end{aligned} \tag{2}$$

Steht dieser Wert für eine kleine oder große Abweichung der Beobachtung von H_0? Diese Frage lässt sich objektiv beantworten, da, gegeben H_0, die approximative Verteilung der Teststatistik bekannt ist und sich die Wahrscheinlichkeit dafür, einen Wert zu beobachten, der mindestens so groß wie der Prüfwert ist, somit unmittelbar aus der Verteilungsfunktion ergibt. Für unseren Prüfwert 3.8 ergibt sich eine Wahrscheinlichkeit von 5.1%; diese bedingte Wahrscheinlichkeit wird als p-Wert bezeichnet.

Der letzte Schritt zur Testentscheidung ist, den p-Wert mit dem *vorher* festgelegten Signifikanzniveau α zu vergleichen. Dabei gilt die folgende Entscheidungsregel:

$p > \alpha \Rightarrow H_0$ wird *nicht* verworfen.

$p \leq \alpha \Rightarrow H_0$ wird verworfen.

Die Standardwahl für α ist 0.05.[11] Damit ergibt sich in unserem Beispiel, dass H_0 auf einem Signifikanzniveau von $\alpha = 0.05$ nicht abgelehnt werden kann. So imposant der positive Effekt von D auf den ersten Blick erschien, er ist statistisch nicht signifikant.

Hieraus folgt jedoch *nicht*, dass H_0 wahr ist, d. h., die angezeigte Version der Webseite wirklich keinen Einfluss auf das Konvertierungsverhalten der Besucher hat. Die Tatsache, dass H_0 nicht verworfen werden kann, bedeutet nur, dass man nicht zeigen konnte, dass H_0 falsch ist (Cohen 1990: 1308). Darum empfiehlt es sich auch nicht, davon zu sprechen, dass H_0 angenommen wurde.

Falsch ist auch die Vorstellung, der p-Wert sei die Wahrscheinlichkeit, mit der H_0 wahr ist. Damit würde er nämlich zugleich eine Wahrscheinlichkeit dafür angeben, dass H_0 falsch ist, was widersprüchlich wäre, da er ja gerade unter der Annahme bestimmt wurde, dass H_0 wahr ist (Goodman 2008: 135).

Wenn H_0 wie hier nicht abgelehnt wird, stellt sich die Frage, ob H_0 tatsächlich wahr und die Testentscheidung damit korrekt ist, oder ob H_0 zu Unrecht nicht verworfen wurde und ein β-Fehler begangen wurde. Wie sich β ermitteln lässt, von welchen Parametern es abhängt und welche Werte akzeptabel sind, dies werde ich in Abschnitt 5 erläutern.

4 Effektstärke

> *The primary product of a research inquiry is one or more measures of effect size, not P values.*
>
> —Jacob Cohen (1990: 1310)

Wie stellt sich die Lage dar, wenn H_0 abgelehnt wird? Betrachten wir dazu erneut die fiktiven Ergebnisse in Tabelle 1, aber dieses Mal nehmen wir an $b = 38$. Damit ergibt sich ein p-Wert von 0.041 und gemäß der obigen Entscheidungsregel wird H_0 folglich auf einem Signifikanzniveau von $\alpha = 0.05$

[11] Andere gängige Werte sind 0.1 und 0.01. Allen drei gemein ist ihre Willkürlichkeit (Stigler 2008: 12).

abgelehnt. Die Idee dahinter ist: Wenn die Wahrscheinlichkeit, die Häufigkeiten *a*–*d* unter H_0 zu beobachten, geringer als 5% ist, dann scheint es plausibel, H_0 zu verwerfen und sich stattdessen für H_1 zu entscheiden.

Die Tatsache, dass H_0 abgelehnt wurde, sollte allerdings nicht zu der Annahme verführen, dass der beobachtete Unterschied zwischen den beiden KR in irgendeiner Hinsicht relevant ist. Hätte man nämlich die ursprünglichen Häufigkeiten aus Tabelle 1 je um den Faktor 10 erhöht, hätte sich ein p-Wert von 0.000 ergeben. Obwohl die Differenz zwischen den KR in beiden Fällen gleich ist, wurde sie nur im zweiten Fall als statistisch signifikant eingestuft. Daraus folgt: Der p-Wert zeigt zwar an, ob ein Effekt existiert, taugt aber offensichtlich nicht als Indikator der eigentlichen Effektstärke.

4.1 Effektmaße

Hierfür wesentlich besser geeignet sind die im Eingangszitat erwähnten Effektmaße, die anzeigen, wie stark ein bestimmter Effekt in der Population ist. Da die zu ihrer Berechnung nötigen Populationsparameter in der Praxis meistens unbekannt sind, werden typischerweise Stichprobeneffekte berechnet. Diese fungieren dann als (verzerrte) Punktschätzer des wahren Populationseffekts und werden darum üblicherweise mit einem Hut gekennzeichnet. Die Unterscheidung zwischen Stichproben- und Populationseffekten ist unabdingbar für die im nächsten Kapitel zu behandelnde Teststärkeanalyse.

Effektmaße sind so konstruiert, dass ihr Wert 0 ist, wenn H_0 wahr ist, und umso größer wird, je mehr die Population H_0 widerspricht. Im Gebrauch sind sowohl standardisierte wie nichtstandardisierte Maße. Erstere sind absolute Größen. Sie treten z. B. beim Vergleich von zwei Populationsmittelwerten auf und sind meist direkt interpretierbar.[12] Letztere sind dimensionslos, weshalb für ihre Bewertung Richtwerte nötig sind.

4.2 Cohens w

Für Kontingenztafeln hat Cohen (1988: 216) das standardisierte Maß w vorgeschlagen. Es ist definiert als

[12]Ein Webshopbetreiber könnte z. B. untersuchen, ob Frauen oder Männer im Durchschnitt mehr Geld pro Einkauf bei ihm ausgeben. Der Unterschied in Euro wäre dann ein unstandardisiertes Effektmaß für den Effekt des Geschlechts auf den Einkaufswert.

$$w = \sqrt{\sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}}, \tag{3}$$

wobei p_{0i} (p_{1i}) die relative Häufigkeit der i-ten Zelle unter H_0 (H_1) und k die Anzahl der Zellen ist. Augenfällig ist die strukturelle Ähnlichkeit zwischen w und χ^2. Inhaltlich gibt es allerdings zwei Unterschiede: w ist für Populationswerte definiert und basiert auf relativen statt absoluten Häufigkeiten. Erstere ergeben sich, indem man letztere durch den Stichprobenumfang teilt. Dadurch wird erreicht, dass w nicht mehr von der Stichprobengröße abhängt.

Einfluss auf w haben stattdessen das absolute Niveau der KR in der Kontrollgruppe und der prozentuale Effekt der Designänderung. Dieser Zusammenhang ist in Tablle 3 dargestellt.[13] Demnach steigt w mit der KR in der Kontrollgruppe und dem Effekt der Designänderung. Anders als man zunächst meinen könnte, bewirkt also eine KR-Verbesserung (oder -Verschlechterung) von z. B. 10% *nicht* stets die gleiche Effektstärke gemessen in w.

Mit Tabelle 3 lässt sich $\hat{w}$ für das Beispiel aus Tabelle 1 abschätzen. Dort lag die KR in der Kontrollgruppe bei 18.9%; die prozentuale Differenz zwischen den KR betrug 55.6%. Damit schätzt man $0.11 \leq \hat{w} \leq 0.13$, wobei der exakte Wert (auf zwei Kommastellen gerundet) 0.12 beträgt. Interessant dabei ist: Mit $b = 38$ ergibt sich eine nur geringfügig bessere Effektstärke von $\hat{w} = 0.13$. Dies zeigt erneut, dass es falsch wäre, die Analyse nach der Berechnung des p-Werts als beendet abzubrechen und etwaige *statistische* Signifikanz mit inhaltlicher gleichzusetzen.

Wie aber sind diese Ergebnisse qualitativ zu bewerten? Nach Cohen (1992: 157) entspricht $w = 0.1$ einem schwachen Effekt, $w = 0.3$ einem mittleren und $w = 0.5$ einem starken. Demnach stehen beide Werte für einen schwachen Effekt der Designänderung D auf das Konvertierungsverhalten der Besucher. Cohen selbst räumt allerdings ein, dass die beste Richtschnur zur Übersetzung von w-Werten in qualitative Begriffe („schwach“, „mittel“, „stark“ etc.) kontextspezifische Erfahrungswerte sind (Cohen 1988: 224). Auch Erdfelder et al. (2010: 364) halten die „unreflektierte Verwendung einheitlicher Normen für Effektgrößen über Anwendungskontexte hinweg“ für „problematisch“. Doch es gibt noch eine weitere Klippe bei der Bewertung

[13] Die Tabellenwerte beruhen auf der (bei technisch korrekt implementierten A/B-Tests annähernd erfüllten) Annahme, dass beide Gruppen exakt gleich groß sind. Bei ungleichen Gruppengrößen ergeben sich abweichende Werte, wobei der grundlegende Zusammenhang aber bestehen bleibt.

Tabelle 3: w-Werte

Effekt (%)	**KR in Kontrollgruppe (%)**											
	1	2.5	5	10	20	30	40	50	60	70	80	90
1	0.001	0.001	0.001	0.002	0.002	0.003	0.004	0.005	0.006	0.008	0.010	0.015
2	0.001	0.002	0.002	0.003	0.005	0.007	0.008	0.010	0.012	0.015	0.020	0.031
3	0.001	0.002	0.003	0.005	0.007	0.010	0.012	0.015	0.018	0.023	0.031	0.048
4	0.002	0.003	0.005	0.007	0.010	0.013	0.016	0.020	0.025	0.031	0.041	0.066
5	0.002	0.004	0.006	0.008	0.012	0.016	0.020	0.025	0.031	0.039	0.052	0.084
10	0.005	0.008	0.011	0.016	0.025	0.032	0.041	0.050	0.062	0.079	0.109	0.197
20	0.010	0.015	0.022	0.032	0.048	0.064	0.081	0.101	0.127	0.166	0.246	0.905
30	0.014	0.022	0.032	0.047	0.071	0.095	0.120	0.152	0.195	0.265	0.442	-
40	0.018	0.029	0.042	0.062	0.094	0.125	0.160	0.204	0.267	0.382	0.816	-
50	0.023	0.036	0.052	0.076	0.115	0.155	0.200	0.258	0.346	0.529	-	-
60	0.026	0.042	0.061	0.089	0.137	0.185	0.240	0.314	0.453	0.734	-	-
70	0.030	0.048	0.070	0.102	0.158	0.214	0.281	0.374	0.535	1.075	-	-
80	0.034	0.054	0.078	0.115	0.178	0.243	0.322	0.436	0.655	2.000	-	-
90	0.038	0.06	0.087	0.128	0.198	0.272	0.365	0.504	0.803	-	-	-
100	0.014	0.066	0.032	0.047	0.071	0.095	0.120	0.152	1.000	-	-	-

von $\hat{w}$: Wie ich in Abschnitt 5.2.1 zeigen werde, ist $\hat{w}$ in der Praxis zumeist *kein* erwartungstreuer Schätzer von w.

4.3 Konfidenzintervall um w

Punktschätzer wie $\hat{w}$ haben den Nachteil, dass sie uns nichts über die Genauigkeit der Schätzung im Einzelfall verraten. Sind sie erwartungstreu, so wissen wir nur, dass sie *im Mittel* den wahren Wert treffen. Ebenso taugt ihre Standardabweichung nur als generelles Maß der Abweichung. Abhilfe schaffen hier Konfidenzintervalle zum Konfidenzniveau $1 - \alpha$.

Interpretiert werden solche Intervalle wie folgt: Würde man das Experiment hinter einer gegebenen Stichprobe unendlich oft wiederholen und jeweils ein Konfidenzintervall zum Niveau $1-\alpha$ berechnen, der Anteil der Intervalle, die den wahren Wert überdecken, würde $1-\alpha$ betragen. Zu denken, dass ein einzelnes Konfidenzintervall jenen Bereich angibt, in dem der wahre Wert mit der Wahrscheinlichkeit $1-\alpha$ liegt, ist ein weit verbreitetes Missverständnis.[14]

Zur Berechnung von Konfidenzintervallen um Effektmaße gibt es zwei Methoden: Bei der gängigeren berechnet man ein Konfidenzintervall um den Nichtzentralitätsparameter der entsprechenden nichtzentralen Verteilung (Cumming & Finch 2001). Die Alternative hierzu ist mittels Bootstrapping eine empirische Verteilung des Effektmaßes zu generieren.[15] Die Frage, welche Methode zu bevorzugen ist, wurde nach meiner Kenntnis für Cohens w bisher noch nicht untersucht; für Cohens d, einem Effektmaß für Mittelwertunterschiede, dahingegen schon (Kelley 2005, Chen & Peng 2014). Dabei gibt es allerdings (noch) keinen klaren „Sieger“ und selbst wenn es ihn gäbe, wäre fraglich, inwiefern sich die Ergebnisse von einem Maß auf das andere übertragen lassen. Ich wähle darum hier den Standardweg über den Nichtzentralitätsparameter λ der nichtzentralen Chi-Quadrat-Verteilung, der den Vorteil hat, dass man, wie sich gleich zeigen wird, schon *vor* Berechnen des 95%-Konfidenzintervalls sieht, ob es den Wert 0 beinhalten wird oder nicht.

In Abschnitt 3.2 habe ich dargelegt, dass die Teststatistik des CQUT unter H_0 approximativ Chi-Quadrat verteilt ist. Unter H_1 dagegen folgt χ^2 approximativ einer nichtzentralen Chi-Quadrat-Verteilung, deren Lage und

[14] Ein solcher Überdeckungsbereich, häufig Kredibilitätsintervall genannt, kann im Rahmen der frequentistischen Inferenzstatistik nicht ermittelt werden. Hierfür müsste ein bayesianischer Ansatz gewählt werden.

[15] Banjanovic & Osborne (2016) demonstrieren diesen Ansatz für verschiedene Effektmaße und Konfidenzintervalltypen.

Form durch $\lambda > 0$ gegeben ist.[16] Für die untere (obere) Grenze des Konfidenzintervalls um λ sucht man nun jene nichtzentrale Verteilung, die mit Wahrscheinlichkeit $1-\frac{\alpha}{2}$ ($\frac{\alpha}{2}$) einen Wert kleiner gleich der beobachteten Teststatistik (Prüfwert) generiert. Man sucht also zwei λ-Werte, so dass gilt:

$$P(\chi^2_{\lambda_u} \leq \chi^2_{\text{beob.}}) = 1 - \frac{\alpha}{2} \tag{4}$$

$$P(\chi^2_{\lambda_o} \leq \chi^2_{\text{beob.}}) = \frac{\alpha}{2} \tag{5}$$

$$P(\lambda_u \leq \lambda \leq \lambda_o) = 1 - \alpha \tag{6}$$

Allerdings können die Gleichungen 4 und 5 nicht direkt, sondern nur iterativ gelöst werden und dies auch nicht in allen Fällen. Betrachten wir dazu noch einmal das obige Beispiel aus Tabelle 1 mit dem Prüfwert 3.8. Wie lautet in diesem Fall das 95%-Schätzintervall um w? Zuerst suchen wir eine nichtzentrale Chi-Quadrat Verteilung mit einem Freiheitsgrad, für die gilt, dass 97.5% ihrer Werte nicht größer als 3.8 sind. Es stellt sich heraus, dass eine solche Verteilung nicht existiert, denn selbst für $\lambda = 0$ – die gewöhnliche Chi-Quadrat-Verteilung – sind nur 94.9% der Werte kleiner oder gleich 3.8. Erst ab einem Prüfwert von 5.1 lässt sich ein geeignetes λ_u finden. Mit anderen Worten: Ist der Prüfwert wie hier kleiner als 5.1, gilt stets $\lambda_u = 0$. Für λ_o finden wir den Wert 15.3.

Der letzte Schritt ist jetzt, die entsprechenden w-Werte für λ_u und λ_o zu finden. Dafür nutzen wir den Zusammenhang (Cohen 1988: 549):

$$\lambda = w^2 \cdot n\,. \tag{7}$$

Mit $n = 253$ erhalten wir damit das 95%-Schätzintervall [0, 0.25]. Dass es den Wert 0 beinhaltet, bedeutet, dass ein Hypothesentest die Nullhypothese $w = 0$ – D hat keinen Effekt auf die KR – auf dem Signifikanzniveau $\alpha = 0.05$ nicht verwerfen würde (Steiger & Fouladi 1997: 226). Zu dem gleichen Fazit war auch der CQUT gelangt. Doch diese Übereinstimmung ist nicht zwingend. In der mit $b = 38$ modifzierten Version des Beispiels ergibt sich das gleiche 95%-Schätzerintervall, wobei der CQUT dort H_0 auf dem Niveau $\alpha = 0.05$ verworfen hat.

[16] Eine ausführlichere Darstellung der Thematik erfolgt in Abschnitt 5.1.

5 Teststärke

> *When I finally stumbled onto power analysis [...], it was as if I had died and gone to heaven.*
>
> —Jacob Cohen (1990: 1308)

Als Teststärke bezeichnet man die Wahrscheinlichkeit, mit der ein statistischer Test zugunsten einer konkreten H_1 entscheidet, gegeben dass diese richtig ist. Angestrebt werden typischerweise Werte zwischen 80% und 95%. Ein starker Test zeichnet sich dadurch aus, dass er schon geringe Abweichungen von H_0 verlässlich aufdeckt. Umgekehrt wird ein schwacher Test eine Abweichung erst dann als statistisch signifikant ausweisen, wenn sie sehr deutlich ausfällt (Cohen 1988: 56).

5.1 Nichtzentrale Verteilung

Zur Bestimmung der Teststärke benötigt man die Verteilung der Teststatistik unter H_1. Diese heißt nichtzentrale Verteilung und ist eindeutig bestimmt, sofern man einen konkreten Populationseffekt annimmt. Ihr Name rührt daher, dass sie gegenüber der Verteilung unter H_0 verschoben ist. Lage und Form der nichtzentralen Verteilung werden durch den Nichtzentralitätsparameter festgelegt, welcher sich in der Regel als Produkt aus dem Stichprobenumfang und einem standardisierten Effektmaß schreiben lässt – so auch hier, wie wir in Gleichung 7 gesehen haben.

In Abbildung 2 ist die approximative Verteilung von χ^2, der Teststatistik des CQUT, schematisch dargestellt, sowohl unter H_0 (Chi-Quadrat-Verteilung) als auch unter H_1 (nichtzentrale Chi-Quadrat-Verteilung). Zu sehen sind außerdem die Wahrscheinlichkeiten von α- und β-Fehler. Die Abbildung verdeutlicht zwei Dinge: Erstens, die Teststärke entspricht der Fläche unter der nichtzentralen Verteilung, die im Ablehnungsbereich von H_0 liegt. Zweitens, je weiter der kritische Wert nach rechts rückt, je kleiner also α, desto größer β. Mit anderen Worten: Je größer man die Sicherheit wählt, H_0 nicht zu Unrecht zu verwerfen, desto eher läuft man Gefahr, fälschlicherweise an H_0 festzuhalten.

Neben α beeinflusst λ die Teststärke des CQUT. Aus Abbildung 3 folgt, dass bei fixiertem α die Teststärke mit λ steigt und zwar umso schneller, je größer α. Für $\alpha = 0.05$ ergibt sich eine Teststärke von 80% bei $\lambda = 7.9$. Lässt sich dieser Wert in der A/B-Test-Praxis erreichen?

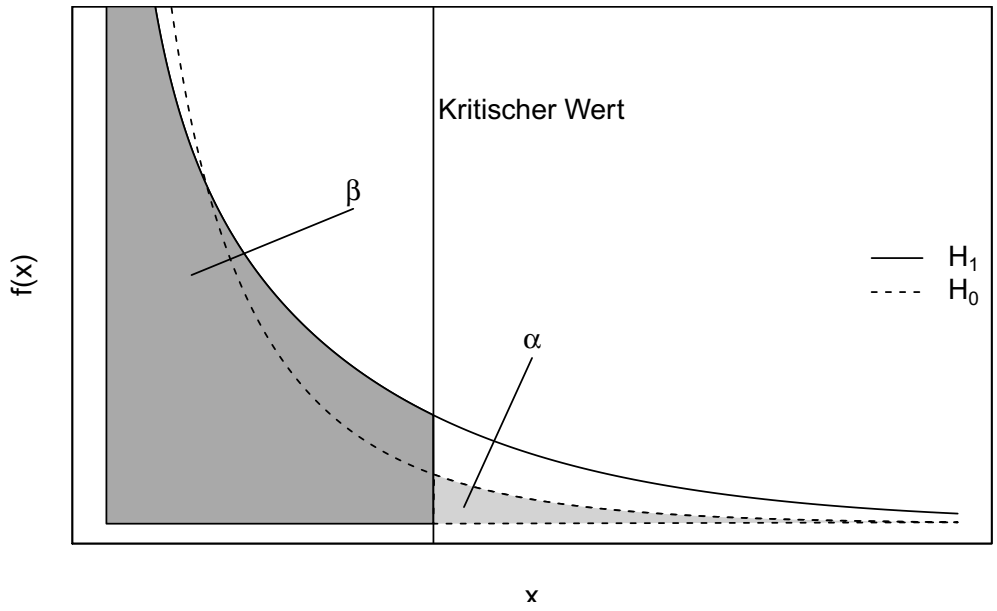

Abbildung 2: Schematische Darstellung der approximativen Verteilung von χ^2 unter H_0 und H_1

Die KR von Landingpages liegt im globalen Durchschnitt einer jüngsten Untersuchung zu Folge bei gerade einmal knapp 2.5% (Monetate 2017). Auf diesem Niveau kann ein λ-Wert von 7.9 nur durch hohe Stichprobenumfänge gepaart mit großen Populationseffekten erzielt werden, wie die folgende Beispielrechnung zeigt: Der Betreiber eines Webshops möchte seine Landingpage, die aktuell eine KR von 2.5% hat, durch eine Designänderung D erhöhen und deren Wirksamkeit in einem A/B-Test überprüfen. Aufgrund der Ergebnisse vergangener A/B-Tests verspricht er sich von D eine KR-Verbesserung in Höhe von 10%, was bei einer bereits optimierten Landingpage ein durchaus ambitioniertes Ziel ist. Damit ergibt sich gemäß Tabelle 3 eine Effektstärke von $w = 0.008$. Um nun auf $\lambda = 7.9$ zu gelangen, ist nach Gleichung 7 ein Stichprobenumfang von $n = 123\,438$ nötig. Würde der Betreiber D eine KR-Verbesserung von „nur“ 5% zutrauen, wäre gar eine viermal so große Stichprobe von Nöten, um die gewünschte Teststärke von 80% zu erreichen.

Derartige Stichprobenumfänge mögen auf den Websites der großen Internetfirmen in akzeptabler Zeit erreichbar sein, für kleinere und mittelgroße Websites dagegen bedeuten sie Testlaufzeiten von mehreren Monaten. Das

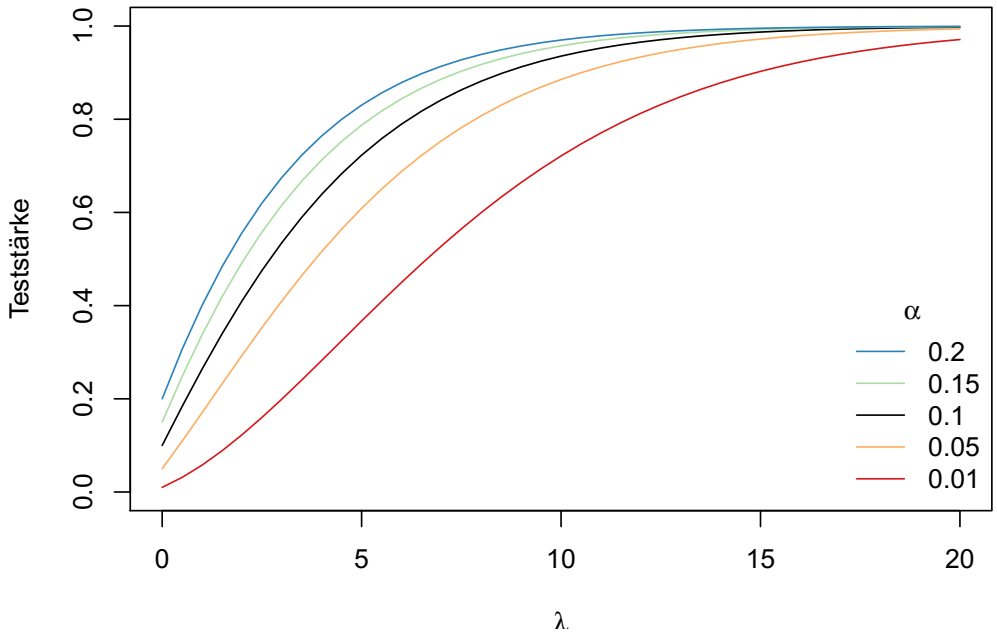

Abbildung 3: Zusammenhang zwischen λ und Teststärke bei $\alpha = 0.05$

Testen konkurrierender Designideen innerhalb kurzer Zeit und damit unter gleichen Marktbedingungen ist so nicht möglich. Die Versuchung liegt also nahe, zugunsten einer kürzeren Testlaufzeit Abstriche bei der Teststärke zu machen. Warum dies keine gute Idee ist und welche Strategie sich stattdessen anbietet, werde ich im Folgenden erläutern.

5.2 Zwei versteckte Nachteile schwacher Tests

Eine geringe Teststärke bedeutet per Definition, dass die Wahrscheinlichkeit, einen existierenden Effekt zu erkennen, gering ist. Wer einen A/B-Test so plant, dass die Teststärke des CQUT bei der nachfolgenden Auswertung gering ist, riskiert also, dass Designänderungen, die bei ihrer Implementierung zu einer KR-Erhöhung führen würden, zu Unrecht als unwirksam aussortiert werden. Mit schwachen Tests wird somit Optimierungspotenzial verschenkt. Doch es gibt noch zwei weitere, weniger offensichtliche Nachteile einer geringen Teststärke.

Tabelle 4: Relative Häufigkeiten (Populationswerte) eines fiktiven A/B-Tests

	A	B	
Konvertiert	0.2	0.3	0.5
Nicht konvertiert	0.3	0.2	0.5
	0.5	0.5	1

5.2.1 Überschätzte Effektstärke

Die gängige Praxis, das Erkennen eines Effekts daran zu knüpfen, dass der p-Wert kleiner als ein bestimmtes α ist, bewirkt, dass Tests mit geringer Teststärke dazu tendieren, den Populationseffekt zu überschätzen (Ioannidis 2008). Neben dem offensichtlichen Nachteil, dass dadurch der Nutzen von D überbewertet wird, gibt es noch einen weiteren: Wie wir in Abschnitt 5.3 sehen werden, benötigt man für die Teststärkeanalyse von A/B-Tests eine Schätzung des Populationseffekts von D. Hierfür bieten sich die beobachteten Effekte vergangener Tests an. Sind diese jedoch im Mittel zu groß, dann führt eine auf ihnen basierende Testplanung dazu, dass die nachfolgenden Tests eine geringere Teststärke haben als erwartet.

Das Problem der überschätzten Effektstärke lässt sich gut anhand einer Simulation illustrieren.[17] Ausgangspunkt ist wieder ein fiktiver A/B-Test. Doch im Unterschied zum letzten Beispiel (Tabelle 1), wo uns Stichprobendaten gegeben waren, nehmen wir dieses Mal an, dass uns die relativen Häufigkeiten *in der Population* gemäß Tabelle 4 bekannt sind. Damit ergibt sich ein Populationseffekt von $w = 0.2$.

Auf Basis der Randverteilungen in Tabelle 4 wurden nun zwei Serien von je 10 000 2×2-Kontingenztafeln (Bootstrap-Stichproben) gezogen. Für die erste Serie wurde der Stichprobenumfang mit $n_1 = 263$ so gewählt, dass der CQUT bei der Auswertung der Stichproben eine Teststärke von 90% hat, gegeben $\alpha = 0.05$ und $w = 0.2$. Bei der zweiten Serie dagegen wurde der Stichprobenumfang auf $n_2 = 51$ verringert, so dass die Teststärke nur noch 30% beträgt. Anschließend wurde für alle Stichproben der p-Wert und Cohens w ermittelt.

Abbildung 4 zeigt für beide Serien ein Histogramm der Effektstärke. Zusätzlich eingezeichnet sind der echte Populationseffekt, der Median aller Ef-

[17] Eine ähnliche Simulation findet sich in Ioannidis (2008).

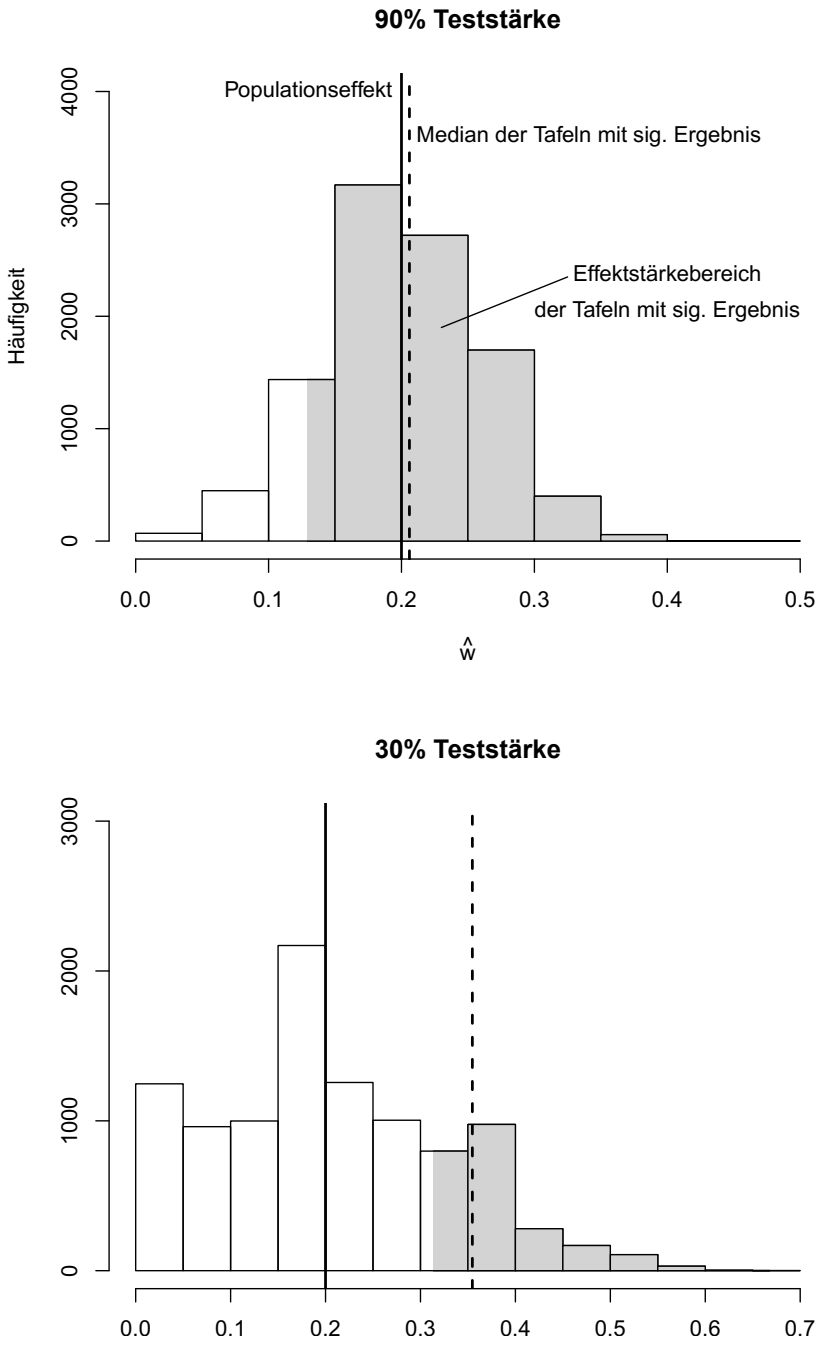

Abbildung 4: Histogramme der beobachteten Effektstärke

Tabelle 5: Lage- und Streuungsparameter der beobachteten Effektstärke

	$\hat{w}$	
	90%	30%
Mittelwert	0.2	0.21
$p \leq 0.05$	0.21	0.38
Median	0.20	0.20
	0.21	0.36
Spannweite	0.47 (0–0.47)	0.69 (0–0.69)
	0.34 (0.13–0.47)	0.38 (0.31–0.69)
IQA	0.08 (0.16–0.24)	0.16 (0.12–0.28)
	0.07 (0.17–0.24)	0.08 (0.32–0.40)
ESD	0.06	0.12
	0.05	0.06

fektstärkewerte von Tafeln mit einem signifikanten Testergebnis und jener Bereich, in den diese Werte gefallen sind. Besonders augenfällig ist, dass dieser Bereich bei den Tafeln, die mit 30% Teststärke ausgewertet wurden, deutlich rechts vom wahren Populationseffekt liegt. Auf die Praxis übertragen bedeutet dies: Hätte man 51 Besucher aus der in Tabelle 4 gezeigten Population gezogen, mit ihnen einen A/B-Test durchgeführt und diesen schließlich mit dem CQUT ausgewertet, man hätte den wahren Populationseffekt deutlich überschätzt. Gemäß Tabelle 5 ist der Median der 30%-Stichproben mit signifikantem Testergebnis ($n = 2361$) ganze 80% größer als der wahre Wert 0.2. Die größte gemessene Effektstärke liegt gar 245% darüber.

Ganz anders stellt sich dagegen die Lage bei den 90%-Stichproben dar. Der Median der Tafeln mit signifikantem Testergebnis ($n = 8920$) liegt hier nur 3% über dem Populationseffekt. Zudem fällt auf, dass die Verteilung der Effektsträke, über alle Tafeln betrachtet, stark an eine Normalverteilung erinnert.[18] Demgegenüber wirkt die Effektstärkeverteilung bei den 30%-Stichproben unregelmäßig und erinnert allenfalls entfernt an eine Gleichverteilung.

Dieser Unterschied schlägt sich in der Streuung nieder: Der Interquartilabstand der 90%-Stichproben ist 50% kleiner als jener der 30%-Stichproben; die Spannweite ist 32% kleiner. Betrachtet man allerdings nur die Tafeln mit signifikantem Ergebnis, ist die Streuung in beiden Serien fast gleich.

[18] Auf diese Beobachtung komme ich in Anhang 8.1 zurück.

5.2.2 Positiver Vorhersagewert

Der zweite versteckte Nachteil schwacher Tests ist: Je geringer die Teststärke, desto geringer die Wahrscheinlichkeit, dass einem als signifikant eingestuften Effekt auch tatsächlich ein Populationseffekt zugrunde liegt. Diese Wahrscheinlichkeit wird als positiver Vorhersagewert (PVW) bezeichnet und darf nicht mit dem p-Wert verwechselt werden. Der PVW ist allgemein wie folgt definiert:

$$\text{PVW} = \frac{\text{Anzahl korrekt klassifizierter positiver Testergebnisse}}{\text{Anzahl positiver Testergebnisse}}\,. \tag{8}$$

Diese Definition ist vor allem in der Medizin gebräuchlich, wo der PVW als Gütekriterium diagnostischer Verfahren fungiert. Hat z. B. ein bestimmter HIV-Test einen PVW von 0.1, heißt das: Im Schnitt ist nur einer von zehn Patienten, bei denen dieser Test positiv ausfällt, auch tatsächlich mit dem HI-Virus infiziert. Will man den PVW für einen statistischen Hypothesentest berechnen, ist es sinnvoll, Gleichung 8 wie folgt zu spezifizieren:

$$\text{PVW} = \frac{(1-\beta)\cdot R}{(1-\beta)\cdot R + (1-R)\cdot \alpha}\,, \tag{9}$$

wobei R für die A-priori-Wahrscheinlichkeit einer falschen Nullhypothese steht.[19] Im Kontext von A/B-Tests ist damit der Anteil der wirksamen Designänderungen an allen getesteten gemeint. Dieser Anteil steht bereits vor der Durchführung eines bestimmten A/B-Tests fest und wird im Unterschied zu α und β nicht durch n, sondern durch die Qualität der getesteten Ideen beeinflusst.

Wie α, β und R zusammenspielen, soll das folgende Beispiel veranschaulichen:[20] Angenommen, ein Shopbetreiber weiß aus langjähriger Erfahrung, dass nur etwa jede fünfte getestete Designänderung einen *echten* Effekt auf die KR seiner Kunden hat ($R = 0.2$), und dass der CQUT bei der Auswertung seiner A/B-Tests eine durchschnittliche Teststärke von $1-\beta = 0.3$ hat. Dieser Betreiber führt nun 1 000 A/B-Tests durch, die er mit dem CQUT

[19] Dagegen versteht Ioannidis (2005: 696) unter R das Verhältnis von falschen zu wahren Nullhypothesen in einem bestimmten Forschungsfeld, weshalb er zu einer leicht unterschiedlichen Gleichung für den PVW gelangt.

[20] Ein ganz ähnliches Beispiel findet sich bei Sterne & Smith (2002).

Tabelle 6: Beispiel zum PVW

	H_0 ist wahr	H_0 ist falsch	
H_0 wird nicht verworfen	760	140	900
H_0 wird verworfen	40	60	100
	800	200	1000

auf einem Signifikanzniveau von $\alpha = 0.05$ auswertet. Die erste Annahme bedeutet, dass von den 1 000 Nullhypothesen 200 falsch sind. Von diesen 200 wiederum werden wegen der zweiten Annahme nur 60 zu Recht verworfen. Von den 800 wahren Nullhypothesen werden wegen $\alpha = 0.05$ 40 zu Unrecht verworfen. Insgesamt werden damit 100 Nullhypothesen verworfen, so dass sich ein PVW von 0.6 ergibt.

Aus praktischer Sicht ist ein PVW von 0.6 augenscheinlich unbefriedigend, bedeutet er doch, dass nur etwas mehr als die Hälfte aller als signifikant eingestuften Ergebnisse auch tatsächlich einen wahren Effekt widerspiegeln. Was könnte der Shopbetreiber also tun, um den PVW zu erhöhen? Ein Weg ist, die Fehlerwahrscheinlichkeiten α und β zu verringern. Sofern der jeweils andere Fehler nicht wachsen soll, bedeutet dies allerdings, dass der Stichprobenumfang n vergrößert werden muss.

Sollte dieser Weg z. B. aus Zeitmangel versperrt sein, bleibt nur die Möglichkeit, R zu erhöhen. Gemäß Abbildung 5 gilt für ein gegebenes Teststärkeniveau: Je höher R, desto höher der PVW. Mit anderen Worten: Je größer die A-priori-Wahrscheinlichkeit für einen wahren Effekt, desto größer die Wahrscheinlichkeit, dass ein signifikantes Testergebnis kein „falscher Alarm“ ist. Dies leuchtet intuitiv ein. Aber wie kann R im Kontext von A/B-Tests erhöht werden?

Der direkteste Weg, den Anteil der wirksamen Designänderungen an allen getesteten zu erhöhen, ist offenbar, die Anzahl der insgesamt getesteten Änderungen zu reduzieren und nur die besonders aussichtsreichen zu testen. Diese Vorauswahl kann z. B. durch Usability-Tests oder die verkaufspsychologische Fachliteratur unterstützt werden. Durch letztere weiß man z. B., dass Verbraucher reduzierte Produkte mit größerer Wahrscheinlichkeit kaufen, wenn deren Preise klein und nicht, wie häufig zu sehen, groß gedruckt sind (Coulter & Coulter 2005). Fließen solche Erkenntnisse in A/B-Tests ein, steigt die A-priori-Wahrscheinlichkeit, dass eine getestete Änderung einen

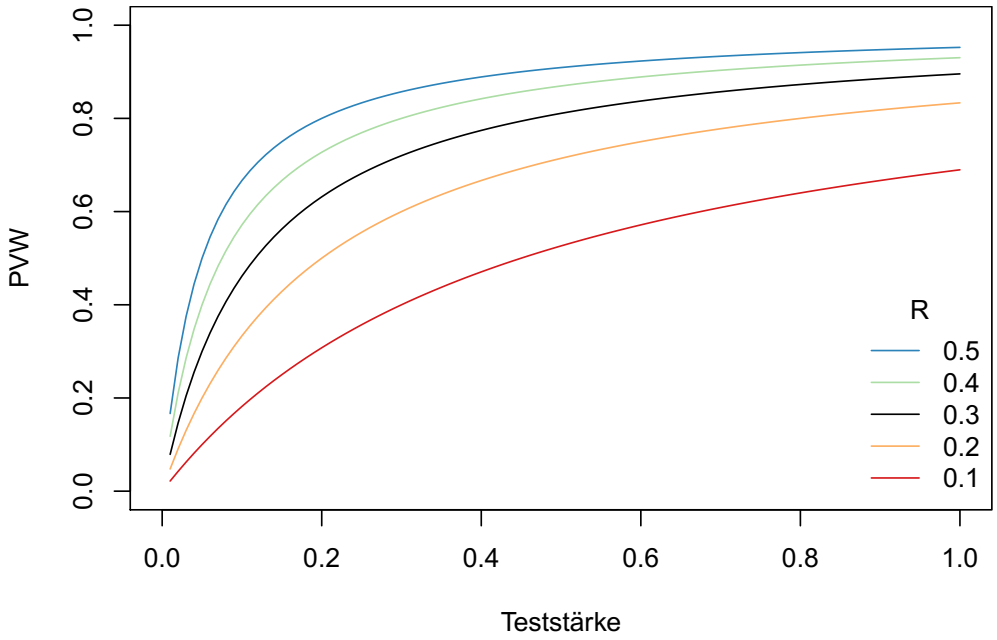

Abbildung 5: Zusammenhang zwischen PVW und Teststärke bei $\alpha = 0.05$

echten Effekt auf das Konvertierungsverhalten der Besucher hat.

5.3 Verschiedene Formen der Teststärkeanalyse

Im letzten Abschnitt habe ich dargelegt, warum es problematisch ist, A/B-Tests so durchzuführen, dass der CQUT bei der Auswertung eine geringe Teststärke hat. Nun soll es näher um die Frage gehen, wie und vor allem wann man die Teststärke eines geplanten A/B-Tests ermittelt. Dazu betrachten wir unterschiedliche Formen der Teststärkeanalyse.

5.3.1 Retrospektive Analyse

Wenn ein Test die Nullhypothese nicht verwirft, scheint es naheliegend, anhand der *beobachteten* Effektstärke die Teststärke zu berechnen. Im Fall einer geringen Teststärke würde die Schlussfolgerung lauten, dass der Test zu schwach war, um den beobachteten Effekt zu entdecken. Das Problem dabei ist: Diese Schlussfolgerung ist unausweichlich, denn Tests mit einem hohen p-Wert haben stets eine geringe beobachtete Teststärke und umgekehrt (Hoe-

Tabelle 7: Lage- und Streuungsparameter der beobachteten Teststärke

	$1 - \hat{\beta}$	
	90%	30%
Mittelwert	0.82	0.36
Median	0.9	0.29
Spannweite	0.95 (0.05 – 1)	0.95 (0.05 – 1)
IQA	0.24 (0.74 – 0.98)	0.41 (0.13 – 0.54)
ESD	0.20	0.25

nig & Heisey 2001). Das Berechnen der beobachteten Teststärke bringt einem Experimentator also keinen Erkenntnisgewinn in Bezug auf die Frage, ob die A-priori-Wahrscheinlichkeit für das Erkennen eines Populationseffekts bei seiner Untersuchung hoch oder niedrig war.

Abbildung 6 zeigt den Zusammenhang zwischen p-Wert und beobachteter Teststärke für die beiden Serien von je 10 000 Bootstrap-Stichproben aus Abschnitt 5.2.1. Demnach hatten in beiden Serien alle Stichproben mit einem nicht signifikanten Testresultat eine beobachtete Teststärke von höchstens 0.5. Ein Anhänger der obigen Logik hätte somit bei allen 90%-Stichproben mit nicht signifikantem Ergebnis ($n = 1080$) zu Unrecht geschlossen, dass sein Test zu schwach zum Erkennen eines Unterschieds in der Höhe des beobachteten Effekts war.

Wie ist es um die Tauglichkeit der beobachteten Testärke als Punktschätzer der wahren Teststärke bestellt? Betrachten wir zunächst die 90%-Stichproben. Gemäß Tabelle 7 liegt der Mittelwert (0.82) hier knapp 9% unter dem wahren Wert. Dass der Median (0.9) größer als der Mittelwert ist und näher am oberen als am unteren Rand des Interquartilabstands liegt, widerspiegelt die im oberen Histogramm in Abbildung 6 deutlich zu sehende Linksschiefe der Verteilung von $1 - \hat{\beta}$. Wenig erfreulich ist aus praktischer Sicht, dass 32% der beobachteten Werte kleiner als 0.8 sind und damit zu Unrecht als zu gering eingestuft worden wären.

Bei den 30%-Stichproben gelingt die Schätzung der wahren Teststärke anhand der beobachteten deutlich schlechter als bei den 90%-Stichproben, in Bezug auf sowohl Genauigkeit als auch Präzision. Der wahre Wert wird hier im Schnitt um 20% *über*schätzt. Die Streuung, gemessen am Interquartilabstand, ist um 71% höher als bei den 90%-Stichproben. Dass die Streuung

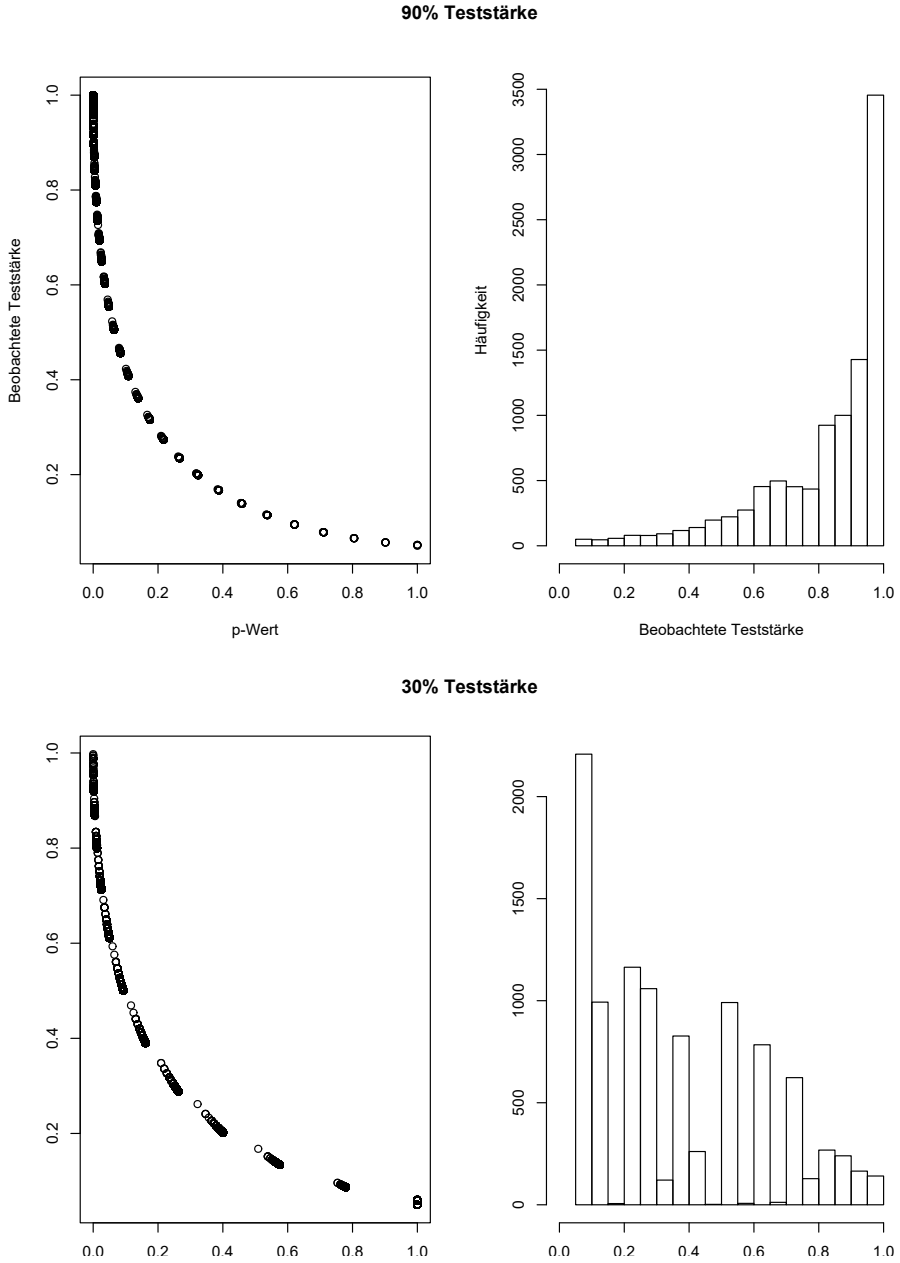

Abbildung 6: Zusammenhang zwischen p-Wert und beobachteter Teststärke sowie Histogramme der beobachteten Teststärke

auch absolut betrachtet hoch ist, zeigt sich daran, dass die empirische Standardabweichung mit 0.25 keine 20% kleiner als der wahre Wert selbst ist. Es wundert daher nicht, dass bei immerhin 813 Stichproben die beobachtete Teststärke gar über 0.8 liegt und diese somit einen Experimentator zu der irrigen Annahme eines ausreichend starken Tests geführt hätte.

Zusammenfassend ergibt sich für die Genauigkeit der Teststärkeschätztung in den beiden Serien das folgende Muster: Ist die wahre Teststärke gering, wird sie durch die beobachtete Teststärke im Mittel überschätzt. Ist sie dagegen hoch, wird sie im Mittel unterschätzt.[21] Da man in der Praxis nicht weiß, mit welchem der beiden Fälle man es zu tun hat, in welche Richtung man also die beobachtete Teststärke korrigieren muss, ist nicht zu erkennen, welchen Nutzen die retrospektive Analyse bei der Auswertung von A/B-Tests mit dem CQUT hat.[22]

5.3.2 A-priori-Analyse

Im Unterschied zur retrospektiven Analyse erfolgt die A-priori-Analyse vor der Testdurchführung. Ihr Ziel ist auch nicht, eine Schätzung der wahren Teststärke abzugeben, sondern zu klären, wie groß n, α und w gewählt werden bzw. mindestens ausfallen müssen, um eine bestimmte Wunschteststärke zu erreichen. Konkret geht es um folgende Fragestellungen:[23]

- **Stichprobenumfangsplanung:** Wie muss n gewählt werden, gegeben eine bestimmte Teststärke, α und w? Wichtig hierbei ist, dass der Experimentator eine fundierte Vorstellung davon hat, welche prozentuale Verbesserung durch die getestete Designänderung D zu erwarten ist. Wird diese zu optimistisch bewertet, ist das gewählte n zu gering, um den vorhandenen Effekt zu entdecken. Wird der Nutzen von D hingegen zu gering angesetzt, läuft der A/B-Test länger als eigentlich nötig. Da das zweite Szenario für die meisten Betreiber langfristig ökonomisch weniger problematisch sein dürfte als das erste, empfiehlt es sich, D

[21] Zu diesem Resultat gelangen auch Yuan & Maxwell (2005) in einer Analyse der beobachteten Teststärke beim Testen auf Mittelwertunterschiede.

[22] Dieses negative Fazit zum Nutzen der beobachteten Teststärke gilt nicht nur für den hier vorliegenden Anwendungsfall. Die retrospektive Teststärkeanalyse hat insgesamt betrachtet keinen guten Ruf in der Literatur. Vgl. z. B. Thomas (1997), Hoenig & Heisey (2001) und Ryan (2013: 9–10), der eine Übersicht kritischer Stimmen gibt.

[23] Die folgende Darstellung orientiert sich lose an Erdfelder et al. (2010).

im Zweifel eher konservativ zu bewerten. Ein Beispiel der Stichprobenumfangsplanung haben wir in Abschnitt 5.1 gesehen, als ich dargelegt habe, warum die Forderung nach einer Teststärke von 80% in der Praxis häufig zu sehr langen A/B-Testlaufzeiten führt.

- **Kriteriumsanalyse:** Wie muss α gewählt werden, gegeben eine bestimmte Teststärke, n und w? Diese Fragestellung tritt auf, wenn n durch äußere Umstände fixiert ist und es aus ökonomischen Gründen angezeigt ist, β und nicht α zu kontrollieren. Dass dies im Rahmen von A/B-Tests häufig der Fall ist, werde ich im nächsten Abschnitt erläutern.

- **Minimum detectable effect (MDE):** Wie groß muss der Populationseffekt von D mindestens sein, um mit einer bestimmten Teststärke gegeben α und n entdeckt zu werden? Der MDE wird primär zum Zweck der Evaluation bereits durchgeführter Hypothesentests ermittelt.

Nutzen und Methodik der A-priori-Analyse wurden erstmals durch Cohen (1969) systematisch beschrieben. Doch entgegen seiner damals geäußerten Hoffnung blieb dies ohne großen Einfluss auf die psychologische Forschungspraxis.[24] Ein wesentlicher Grund hierfür dürfte sein, dass die A-priori-Analyse nicht selten zu unrealisierbar großen Stichprobenumfängen führt, wie Cohen (1990: 1309) einige Jahre später selbst einräumte. Der am häufigsten gewählte Ausweg ist dann, Abstriche bei der Teststärke zu machen oder diese gleich ganz zu ignorieren, so als ob ein β-Fehler *per se* weniger problematisch wäre als ein α-Fehler.

Die bessere Alternative hierzu ist, im Einzelfall zu prüfen, in welchem Verhältnis die negativen Auswirkungen beider Fehler zueinander stehen, und dann dieses Verhältnis als Grundlage zur Stichprobenumfangsplanung zu nehmen. Davon handelt der nächste Abschnitt.

5.3.3 Kompromissanalyse

Der Umststand, dass die Publikation von p-Werten in den meisten Fachzeitschriften ungleich verbreiteter ist als die von Teststärkewerten, verführt

[24] Cohen (1962) hatte gezeigt, dass der Median der Teststärke aller Studien, die im 1960er-Jahrgang des „Journal of Abnormal Pyschology“ veröffentlicht wurden, bei 0.46 lag, einen mittleren Effekt vorausgesetzt. Als Sedlmeier & Gigerenzer (1989) diese Untersuchung anhand des 1984er-Jahrgangs wiederholten, war der Wert sogar leicht gesunken auf 0.44.

leicht zu der Annahme, dass die Kontrolle von α *stets* wichtiger als die von β ist. Doch eine solche Hierarchie existiert nicht, denn welcher Fehler – α- oder β-Fehler – problematischer ist, hängt ganz vom Betrachtungskontext ab. Dass einer problematischer ist als der andere, bedeutet außerdem nicht, dass es ratsam ist, den jeweils anderen überhaupt nicht zu kontrollieren. Es bedeutet vielmehr, dass man sich als Experimentator die Frage stellen sollte, in welchem Verhältnis die Konsequenzen beider Fehler zueinander stehen. Dann kann man nämlich dieses Verhältnis anstatt einzelner Fehlerwahrscheinlichkeiten kontrollieren. Genau darum geht es bei der von Erdfelder (1984) vorgeschlagenen Kompromissanalyse, deren Vorgehensweise wie folgt lautet:

> [M]an legt das maximal vertretbare N sowie die zu entdeckende Devianz W fest und überlegt sich, in welchem Ausmaß α- und β-Fehler unterschiedlichen „Schaden“ anrichten. Diese Überlegungen sollten in der Spezifikation eines Quotienten $q = \beta/\alpha$ münden, der angibt, wieviel Mal größer die β- im Vergleich zur α-Fehlerwahrscheinlichkeit sein darf. Der Quotient $q = 1$ würde indizieren, daß man die Konsequenzen von α- und β-Fehlern für gleichermaßen gravierend erachtet, während $q > 1$ dem α-Fehler und $q < 1$ dem β-Fehler gravierendere Konsequenzen zuschreibt. Für festgelegte Koeffizienten N, W und q lassen sich dann die vernünftigerweise zu verwendenden Irrtumswahrscheinlichkeiten α und β [...] errechnen. (Erdfelder 1984: 27)

Ein Grund, warum sich dieser Ansatz nicht richtig etablieren konnte, dürfte sein, dass hierbei in der Regel solch unkonventionelle Werte für α und β herauskommen, dass ihre Publikation nur schwer zu erreichen sein dürfte. Aus Sicht eines Websitebetreibers stellt dies freilich kein Manko dar. Für ihn zählt die Gewissheit, dass q seine individuelle Kostenbeurteilung der Fehler abbildet und der Stichprobenumfang realisierbar ist.

Wie sollte q im Kontext von A/B-Tests gewählt werden? Betrachten wir dazu den in Tabelle 8 dargestellten „Schaden“ beider Fehler. Die dort aufgelisteten Punkte eint, dass sie sich einer direkten Quantifizierung entziehen, sei es, weil sie stark vom betrachteten D abhängen, oder weil sie ihrer Natur nach schwer zu quantifzieren sind. Welche Kosten beispielsweise bei der vollständigen, aber nutzlosen Implementierung von D anfallen, hängt ganz davon ab, wie tief D in die Architektur der Seite eingreift. Über die Höhe des zusätzlichen Gewinns, der sich bei der nicht erfolgten Implementierung von

Tabelle 8: Konsequenzen beider Fehler im Kontext von A/B-Tests

α-Fehler	β-Fehler
Die vollständige Umstellung auf D (ggf. auch auf weiteren Seiten der Website) verursacht Kosten und Stabilitätsrisiken, denen keine KR-Steigerung gegenübersteht.	D wird nicht implementiert, obwohl dies die KR erhöht hätte.
Änderungen im Stile von D erscheinen sinnvoll, obwohl sie von vorneherein geringe Erfolgsaussichten haben.	Änderungen im Stile von D werden nicht weiter verfolgt, obwohl sie gute Erfolgsaussichten hätten.
Der zukünftige Umsatz wird überschätzt, weil der KR-Effekt von D ausbleibt.	Das Wachstumspotenzial des die Website betreibenden Unternehmens wird unterschätzt.

D eingestellt hätte, kann man dagegen nur spekulieren. Klar ist aber: Der β-Fehler fungiert als Bremse des Gewinnwachstums.

Ein konkretes, allgemeingültiges q lässt sich somit nicht angeben. Allerdings lässt sich zweifellos sagen, dass es aus Sicht des Websitebetreibers falsch wäre, blind der Standardempfehlung zu folgen, den α-Fehler höher zu gewichten als den β-Fehler. Wie wir in Abschnitt 6.4.1 sehen werden, ist es bei leicht zu implementierenden Designänderungen in der Regel sinnvoll, letzteren (deutlich) höher zu gewichten als ersteren.

6 A/B test pro

Ich stelle nun eine Webanwendung zur Planung und Auswertung von A/B-Tests vor, die ich für die Firma Scout24 entwickelt habe: den A/B test pro (ABTP).[25] Mit dem ABTP können die im letzten Kapitel dargestellten Formen der Teststärkeanalyse durchgeführt sowie A/B-Tests auf Basis des CQUT ausgewertet werden. Eingesetzt wird der ABTP im Produktmanagement der Töchterfirmen Immobilien Scout GmbH und AutoScout24 GmbH. Eine zentrale Vorgabge bei der Entwicklung war daher, dass zur Bedienung

[25]Der englische Titel und die englische Benutzeroberfläche des ABTP rühren daher, dass Englisch neben Deutsch offizielle Unternehmenssprache der Scout24 AG ist.

des ABTP (neben einer ausführlichen Einweisung) lediglich Grundkenntnisse der Inferenzstatistik erforderlich sind.

6.1 Wozu eine weitere Webanwendung?

Wer einen A/B-Test ohne den Einsatz spezieller Software auswerten möchte, wird im Internet schnell fündig. Auf zahlreichen Websites lässt sich kostenlos und unkompliziert ermitteln, ob ein bestimmter KR-Unterschied statistisch signifikant ist oder nicht. Wozu, so wird man fragen, braucht es also eine weitere Webanwendung?

- Alle mir bekannten Anwendungen fokussieren sich ausschließlich auf die Auswertung von A/B-Tests und nicht auf deren Planung. Für die verschiedenen Formen der Teststärkeanalyse ist bisher der Einsatz von Statistik-Software nötig, weshalb die Kontrolle des β-Fehlers in der Praxis oft unterlassen wird.

- Die existierenden Anwendungen geben keine Effektmaße aus. Meistens erfährt der Benutzer noch nicht einmal den exakten p-Wert, sondern lediglich, ob der beobachtete Unterschied auf einem bestimmten Signifikanzniveau statistisch signifikant ist oder nicht.

- Bei vielen Anwendungen ist nicht ersichtlich, welcher Hypothesentest verwendet wird. Somit bleibt für den Benutzer unklar, welche Annahmen der Signifikanzberechnung zugrunde liegen und ob diese in seinem konkreten Anwendungsfall überhaupt erfüllt sind.

- Die meisten Anwendungen liefern keine Hinweise zur korrekten Interpretation des p-Werts, so dass Missverständnisse vorprogrammiert sind.[26] Ferner gibt es Anwendungen mit irreführenden Erläuterungen.[27]

[26] Wie weit solche Missverständnisse verbreitet sind, wurde in mehreren Studien anhand verschiedener Personengruppen demonstriert; siehe z. B. Falk & Greenbaum (1995) für israelische Studenten. Dass selbst Lehrkräfte der Statistik (darunter Professoren) deutscher Universitäten betroffen sind, haben Haller & Krauss (2002) gezeigt.

[27] Als Beispiel sei hier die auf dem CQUT beruhende A/B-Test-Anwendung von `www.blitzrechner.de` genannt, wo ein „Konfidenzniveau $> 95\%$“ mit einer „Irrtumswahrscheinlichkeit $< 5\%$“ gleichgesetzt wird. Dem Benutzer wird dadurch suggeriert, die Wahrscheinlichkeit für einen Fehler bei der Auswertung seines A/B-Tests sei kleiner als 5%. Im Rahmen der frequentistischen Inferenzstatistik stehen allerdings nur die *bedingten* Fehlerwahrscheinlichkeiten α und β zur Verfügung.

Zusammenfassend lässt sich sagen: Zur professionellen Planung und Auswertung von A/B-Tests sind die existierenden kostenlosen Webanwendungen nicht geeignet, da ihr Funktionsumfang zu gering ist und das Testresultat allenfalls von Personen mit statistischem Fachwissen richtig gedeutet werden kann.

6.2 R

Der ABTP ist in der freien Programmiersprache R geschrieben, die Anfang der 90er-Jahre als Weiterentwicklung von S entstand und mittlerweile in Wissenschaft und Wirtschaft weit verbreitet ist. Für den Einsatz von R im Unternehmen sprechen u. a. die folgenden Gründe (Wickham 2014: 1-3):

- R ist frei zugänglich. Im Unterschied zu klassischen Statistikprogrammen wie IBM SPSS Statistics oder Stata fallen also keine Lizenzkosten an.
- Der Funktionsumfang von R kann durch zusätzliche Pakete erheblich erweitert werden. Allein das Comprehensive R Archive Network verwaltet aktuell knapp über 12 000 Pakete (R Foundation o. D.).
- R ist fest in der statistischen Forschung verwurzelt. Neue Verfahren sind meist schon nach kurzer Zeit in R verfügbar.
- Es gibt eine umfangreiche R-interne Hilfe. Spezifische Probleme beim Programmieren in R lassen sich in der Regel schnell mit Hilfe der R-Gemeinde der Entwicklerplattform `www.stackoverflow.com` lösen.

Bei größeren R-Projekten werden meistens zusätzliche Pakete benötigt, so auch beim ABTP. Die zwei wichtigsten davon betrachte ich nun näher.

6.2.1 Shiny

Dynamische Webanwendungen werden normalerweise in einem Webframework wie ASP.NET oder Ruby on Rails entwickelt, wofür Kenntnisse in HTML, CSS und Javascript notwendig sind. In R dagegen gibt es mit Shiny (Chang et al. 2017) ein Paket, das die Entwicklung solcher Anwendungen auch ohne diese Kenntnisse erlaubt. Eine Shiny-Anwendung besteht aus zwei Komponenten:

- **UI:** Hier werden Struktur und Design der Benutzeroberfläche festgelegt, z. B. wie viele Buttons eine Anwendung hat und welche Farbe diese haben. Dazu verwendet man verschachtelte R-Funktionen, welche im Hintergrund HTML-Code generieren.
- **Server:** Hier wird festgelegt, was bei einem bestimmten Input passieren soll, z. B. dass ein Histogramm mit 5 Klassen gezeichnet wird, wenn der Benutzer vorher die Zahl 5 in ein entsprechendes Feld eingegeben und anschließend auf „Histogramm erstellen“ geklickt hat. Diese Logik wird in R geschrieben.

Um Dritten eine Shiny-Anwendung über das Internet oder ein Intranet zur Verfügung zu stellen, ist es notwendig, diese auf einem Server zu hosten, auf dem eine aktive R-Session läuft. Damit wird erreicht, dass In- und Output „live“ miteinander verbunden sind und die Benutzeroberfläche, falls vom Entwickler so gewünscht, direkt auf Input reagiert.

6.2.2 pwr

Wer die Teststärke des CQUT in R analysieren will, muss dafür entweder eine eigene Funktion schreiben oder auf ein externes Paket zurückgreifen. Ich habe für den ABTP letzteren Weg gewählt und mich für das pwr-Paket (Champely 2016) bzw. dessen Funktion `pwr.chisq.test` entschieden. Diese Funktion hat erwartungsgemäß fünf Argumente – w, N, α, $1-\beta$ und DOF –, wobei eines der ersten vier mit `NULL` übergeben werden muss. Dieses fehlende Argument wird dann auf Basis der anderen vier berechnet.

Dazu bedient sich `pwr.chisq.test` der R-Funktion `uniroot`, die Nullstellen einer mathematischen Funktion mittels Bisektion ermittelt, einem simplen numerischen Verfahren, das u. a. voraussetzt, dass die Funktion an den Grenzen des Startintervalls, das den Suchbereich festlegt, unterschiedliche Vorzeichen hat. Beim Testen verschiedener Kombinationen von w, N, α und $1 - \beta$ mit $DOF = 1$ hat sich gezeigt, dass diese Annahme bei „extremen“ Parameterwerten verletzt ist, `pwr.chisq.test` also einen Fehler ausgibt.[28]

Weil das Abfangen von Fehlern selbst eine Fehlerquelle ist, habe ich die Funktion `pwr.chisq.test` so modifiziert, dass sie zur Nullstellenfindung `uniroot.all` aus dem Paket rootSolve (Soetaert 2009) verwendet. Diese Funktion ruft zwar ebenfalls `uniroot` auf, ist aber so konstruiert, dass

[28] Ein solcher Fehler lässt sich z. B. mit den Werten $N = 1000$, $\alpha = 0.6$, $1 - \beta = 0.5$ erzeugen.

`pwr.chisq.test` bei extremen Kombinationen nun keinen Fehler mehr zurückgibt, sondern dem gesuchten Parameter den Wert `Inf` zuweist. Dieses Verhalten ist aus Entwicklersicht unproblematisch.

6.3 Dokumentation

Der ABTP hat drei Tabs:

- **Plan:** Dieser Tab dient der Planung von A/B-Tests und beinhaltet vier verschiedene Formen der Teststärkeanalyse: Kompromissanalyse, Stichprobenumfangsplanung, Kriteriumsanalyse und MDE. Bei allen vier Formen wird angenommen, dass die Besuchergruppen A und B gleich groß sind.
- **Evaluate:** Hier kann ein bereits durchgeführter A/B-Test mit dem CQUT ausgewertet werden. Benötigt werden vier Zahlen: die Anzahl der jeweils konvertierten Besucher in A und B sowie die Anzahl der jeweils nicht konvertierten Besucher in A und B.
- **Get help:** Die Hilfe erläutert alle Eingabefelder und bietet eine Orientierung bei der Frage, wann welche Form der Teststärkeanalyse zu wählen ist.

6.3.1 Plan

Der Plan-Tab, die Startseite des ABTP, besteht aus zwei Teilen: einem Input-Panel und einem Output-Bereich. Das Input-Panel hat vier oder fünf Eingabefelder, je nachdem welche Form der Teststärkeanalyse ausgewählt ist, und stets zwei Buttons: ein Button für die Ergebnisse und ein Button, um die Eingabefelder wieder in ihren ursprünglichen Zustand zurückzusetzen. Abbildung 7 zeigt das Panel der voreingestellten Kompromissanalyse.

Sofern alle Felder korrekt befüllt wurden, erscheint im Output-Bereich eine Tabelle, die den Input zusammenfasst. Bei klar irrtümlichen Eingaben, z. B. Buchstaben statt Zahlen, erscheint statt der Input-Tabelle ein entsprechender Hinweis. Klickt der Benutzer auf den Ergebnis-Button, erscheint unterhalb der Input-Tabelle ein Panel, das die Ergebnisse in Textform präsentiert, um deren Interpretation zu erleichtern. Abbildung 8 zeigt Input-Tabelle und Ergebnis-Panel der Kompromissanalyse. Es folgen Erläuterungen zur technischen Umsetzung der vier Analyseformen.

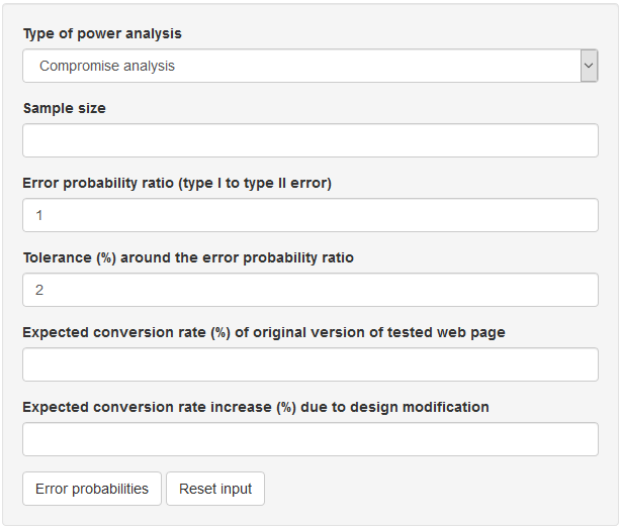

Abbildung 7: Input-Panel Kompromissanalyse

	Your input
Sample size	1000
Error ratio	2
Tolerance (%)	1
CR (%) of original version	30
Expected CR increase (%)	20

If you set the significance level (probability of a type I error) to **31.2** percent and *under the assumption that your design modification has an effect on the conversion rate*, your test will overlook effects greater or equal to 20 percent with a probability of **15.6** percent (probability of a type II error). Smaller effects will be overlooked with a higher probability.

Abbildung 8: Output-Bereich Kompromissanalyse

Kompromissanalyse Input:

- Anzahl der Testteilnehmer
- Verhältnis der Fehlerwahrscheinlichkeiten q
- tolerierte Abweichung (%) von q
- erwartete KR (%) der Kontrollgruppe
- erwartete Verbesserung (%) der KR durch D

Die beiden letztgenannten Größen dienen der Berechnung von Cohens w, das zur Anwendung der Funktion `pwr.chisq.test` erforderlich ist. Auf dieser wiederum basiert der folgende einfache Algorithmus zur Bestimmung der Fehlerwahrscheinlichkeiten α und β: Berechne für $\beta = 0.001$ den zugehörigen α-Wert. Stimmt das Verhältnis beider Werte unter Berücksichtigung der tolerierten Abweichung mit q überein, breche ab. Andernfalls erhöhe β um 0.001 (maximal bis $\beta = 0.999$). Wird auf Anhieb kein passendes Wertepaar gefunden, hilft es in vielen Fällen, die tolerierte Abweichung von q zu erhöhen.

Stichprobenumfangsplanung Input:

- gewünschte Teststärke (%)
- Signifikanzniveau (%)
- erwartete KR (%) der Kontrollgruppe
- erwartete Verbesserung (%) der KR durch D
- erwartete Besucheranzahl pro Woche auf Testseite (optional)

Die Anzahl der Testteilnehmer, die für eine bestimmte Teststärke erforderlich ist, wird durch einen einfachen Aufruf von `pwr.chisq.test` ermittelt. Bei Angabe des erwarteten wöchentlichen Traffics auf die Testseite wird zusätzlich zu dieser Anzahl die entsprechende Testlaufzeit ausgegeben. Dabei gilt: Ist diese kürzer als eine Woche, wird der Benutzer darauf hingewiesen, dass der Test dennoch mindestens eine Woche laufen sollte, um etwaige Unterschiede zwischen den Wochentagen abzubilden. Ist die Laufzeit dagegen

länger als vier Wochen, wird dem Benutzer die Kompromissanalyse nahegelegt. So soll vermieden werden, dass die berechnete Laufzeit vom Benutzer als zu lang abgelehnt wird und dann überhaupt keine Laufzeitplanung stattfindet.

Kriteriumsanalyse Input:

- Anzahl der Testteilnehmer
- gewünschte Teststärke (%)
- erwartete KR (%) der Kontrollgruppe
- erwartete Verbesserung (%) der KR durch D

Die Kriteriumsanalyse basiert ebenfalls auf einem einfachen Aufruf von `pwr.chisq.test`.

MDE Input:

- Anzahl der Testteilnehmer
- gewünschte Teststärke (%)
- Signifikanzniveau (%)
- erwartete KR (%) der Kontrollgruppe

Der MDE in Prozent wird wie folgt ermittelt: Zunächst wird mittels `pwr.chisq.test` w_{MDE} bestimmt. Dann wird Gleichung 3 so umgeschrieben, dass darin statt p_{0i} und p_{1i} die vom Benutzer angegebene erwartete KR der Kontrollgruppe (k) und der gesuchte Effekt durch D (e) vorkommen:[29]

[29] Dies wird in Anhang 8.2 gezeigt.

$$
\begin{aligned}
w^2_{MDE} = & \frac{\Big(0.5 \cdot k - 0.25 \cdot k \cdot (1+e)\Big)^2}{\Big(0.25 \cdot k \cdot (1+e)\Big)} + \\
& \frac{\Big((0.5 - 0.5 \cdot k) - \Big(0.5 - 0.25 \cdot k \cdot (1+e)\Big)\Big)^2}{(0.5 - 0.25 \cdot k \cdot (1+e))} + \\
& \frac{\Big(0.5 \cdot k \cdot e - 0.25 \cdot k \cdot (1+e)\Big)^2}{\Big(0.25 \cdot k \cdot (1+e)\Big)} + \\
& \frac{\Big((0.5 - 0.5 \cdot k \cdot e) - \Big(0.5 - 0.25 \cdot k \cdot (1+e)\Big)\Big)^2}{\Big(0.5 - 0.25 \cdot k \cdot (1+e)\Big)}
\end{aligned}
\tag{10}
$$

Mittels `uniroot.all` ergibt sich nun eine numerische Lösung für e.

6.3.2 Evaluate

Der Evaluate-Tab ist wie der Plan-Tab zweigeteilt: links ein Input-Panel, rechts ein Output-Bereich. Ersteres (Abbildung 9) hat fünf Eingabefelder, vier für die oben genannten Konvertierungshäufigkeiten und eines für das Konfidenzniveau des Konfidenzintervalls um Cohens w. Die Ergebnisse sind in drei Tabellen dargestellt. Per Checkbox können zudem zwei Panels mit Hinweisen zur richtigen Interpretation der Ergebnisse eingeblendet werden. Abbildung 10 zeigt den vollständigen Output-Bereich. Es folgen Erläuterungen zu den drei Tabellen.

Konversionsraten Diese Tabelle enthält drei Zahlen: die KR in A, die KR in B und den Unterschied in Prozent (nicht Pronzentpunkten) zwischen beiden.

Effektstärke Diese Tabelle enthält $\hat{w}$ sowie die obere und untere Grenze des zugehörigen Schätzintervalls. Das zugehörige Textpanel erklärt dem Benutzer, was es bedeutet, wenn ein bestimmter Wert, z. B. 0, nicht im Schätzintervall enthalten ist.

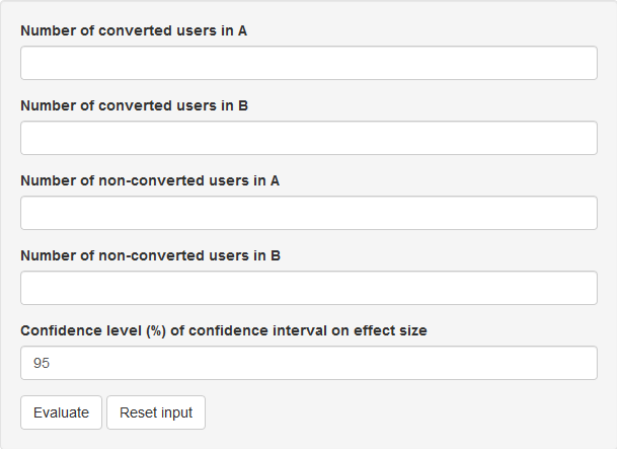

Abbildung 9: Input-Panel Evaluate-Tab

CQUT Diese Tabelle enthält den auf drei Nachkommastellen gerundeten p-Wert des CQUT. Wie dieser korrekt zu interpretieren ist und vor allem, was er nicht leistet, dies erfährt der Benutzer in dem zugehörigen Textpanel.

6.3.3 Get Help

Die Hilfe enthält Erklärungen zu allen Eingabefeldern sowie einen Entscheidungsbaum (Abbildung 11) zur Wahl der im Plan-Tab verfügbaren Analyseformen. Sie kann eine ausführliche Einweisung in den ABTP und ein solides statistisches Grundverständnis der gesamten Materie nicht ersetzen.

6.4 Zwei Anwendungsbeispiele

Zum Abschluss dieses Kapitels demonstriere ich an zwei von Immobilien Scout durchgeführten A/B-Tests, wie der ABTP in der Praxis eingesetzt wird. Bei beiden Tests ging es darum, die KR jener Seite zu steigern, auf der private Immobilieneigentümer zum ersten Mal die (dynamisch errechneten) Anzeigenpreise sehen, wobei mit einer Konversion hier der Klick auf den Button „Produkt wählen“ gemeint ist.[30]

[30]Dadurch gelangt man auf die eigentliche Checkout-Seite.

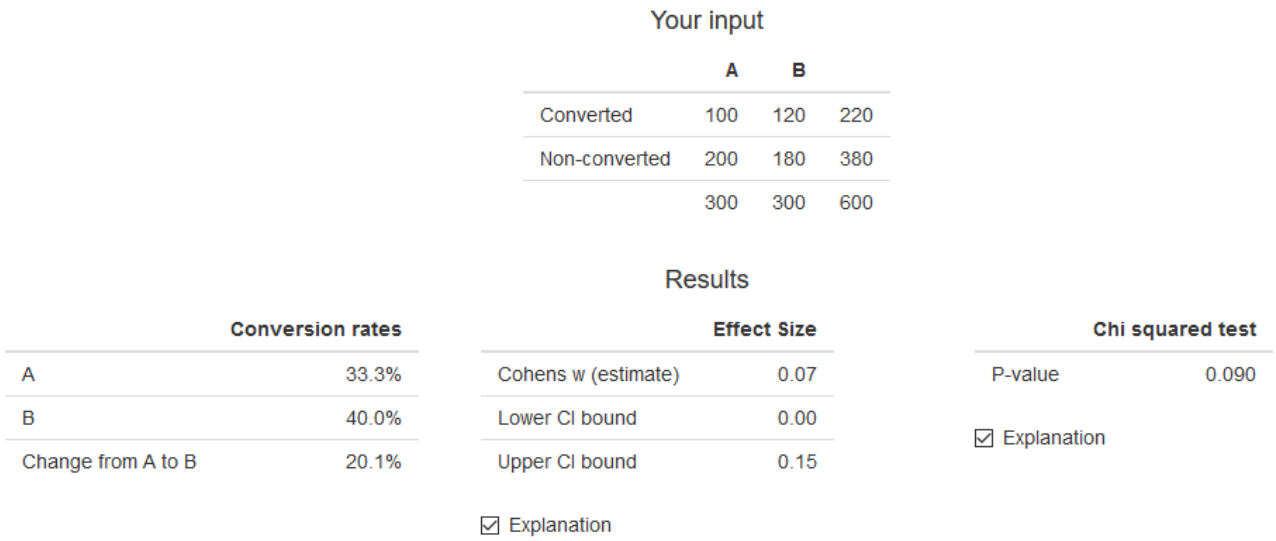

Abbildung 10: Output-Bereich Evaluate-Tab

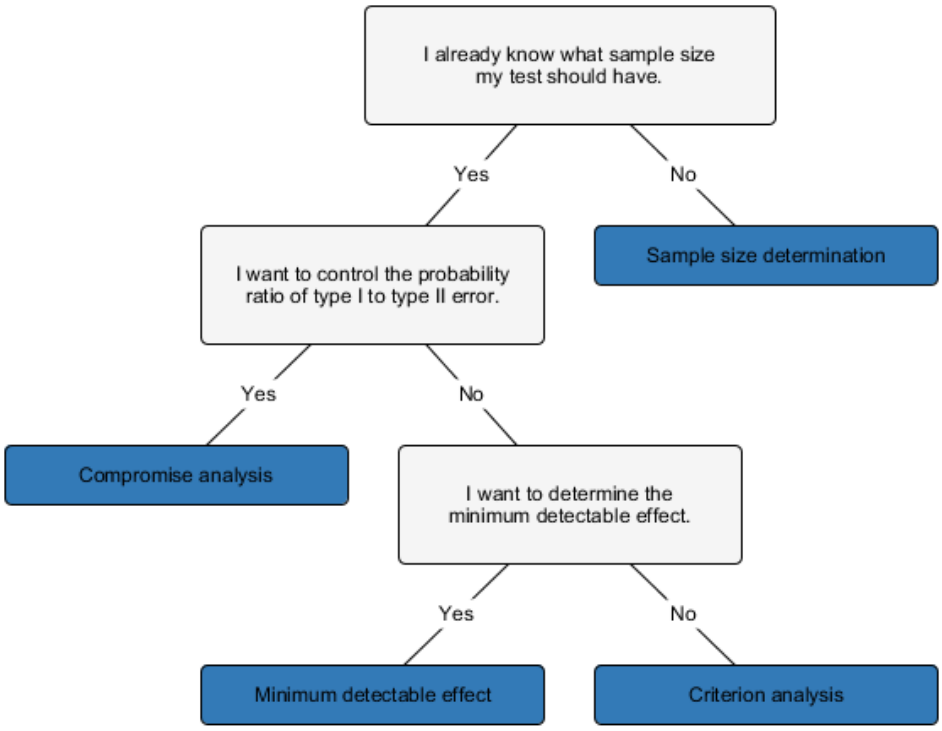

Abbildung 11: Entscheidungsbaum für den Plan-Tab

6.4.1 Steuertipp

Private Immobilieneigentümer können die Kosten einer Immobilienanzeige vollständig von der Steuer absetzen. Da dies vielen Eigentümern nicht bekannt sein dürfte, kam die Idee auf, einen entsprechenden Hinweis direkt bei den Preisen zu platzieren (Abbildung 12), um so die KR zu erhöhen.

Wie wurde dieser Test mit dem ABTP umgesetzt? Zuerst wurde eine Stichprobenumfangsplanung mit den Standardwerten $\alpha = 0.05$ und $\beta = 0.2$ durchgeführt. Die so ermittelte Testlaufzeit lag allerdings deutlich über der ursprünglich angestrebten Laufzeit, bedingt dadurch, dass der erwartete Effekt des Steuertipps mit nur 2% angegeben wurde. Dieser Wert wurde aus vergangenen Tests auf der Seite abgeleitet.

Als Ausweg wurde dann eine Kompromissanalyse mit einem Verhältnis der Fehlerwahrscheinlichkeiten von $3 : 1$ ($\alpha : \beta$) gewählt. Dabei kamen mit $\alpha = 0.75$ und $\beta = 0.25$ zwei Werte heraus, die vom beteiligten Produktteam angesichts der geringen Entwicklungskosten der Testseite als akzeptabel eingestuft wurden. Die deutliche Übergewichtung des α-Fehlers rührte daher, dass im Falle einer KR-Steigerung die vollständige Umstellung auf die Version mit dem Steuertipp nur geringe Implementierungskosten verursacht hätte. Anders gesagt: Das Übersehen eines möglichen Effekts (β-Fehler) wäre hier klar schlechter gewesen als die Implementierung eines Designs, das keinen Einfluss auf die KR hat (α-Fehler).

Der Test ging wie folgt aus: Die KR der Kontrollgruppe war nur marginal schlechter als die KR der Gruppe, die den Steuertipp gesehen hat. H_0 wurde, wie wegen des hohen α-Fehlers zu erwarten war, dennoch abgelehnt, wobei das 95%-Schätzintervall um $\hat{w} = 0.02$ von 0 bis 0.16 reichte.

6.4.2 Startdatum wählen

Wer als Privatperson eine Immobilie auf `www.immobilienscout24.de` inserieren möchte, hat bisher nicht die Möglichkeit, den Startzeitpunkt der Anzeige selbst festzulegen. Es stellt sich daher die Frage, ob eine solche Option die KR steigern würde. Um dies zu testen, wurde neben der Laufzeiteingabe ein Link eingebaut, der einen kleinen Kalender öffnet (Abbildung 14).

Wenn ein Besucher hier ein Datum auswählt und anschließend auf den Button „Produkt wählen“ klickt (Konversion), öffnet sich anders als sonst nicht direkt die Checkout-Seite, sondern zunächst die in Abbildung 13 gezeigte Mitteilung. Der Grund hierfür ist, dass eine tatsächliche Implementie-

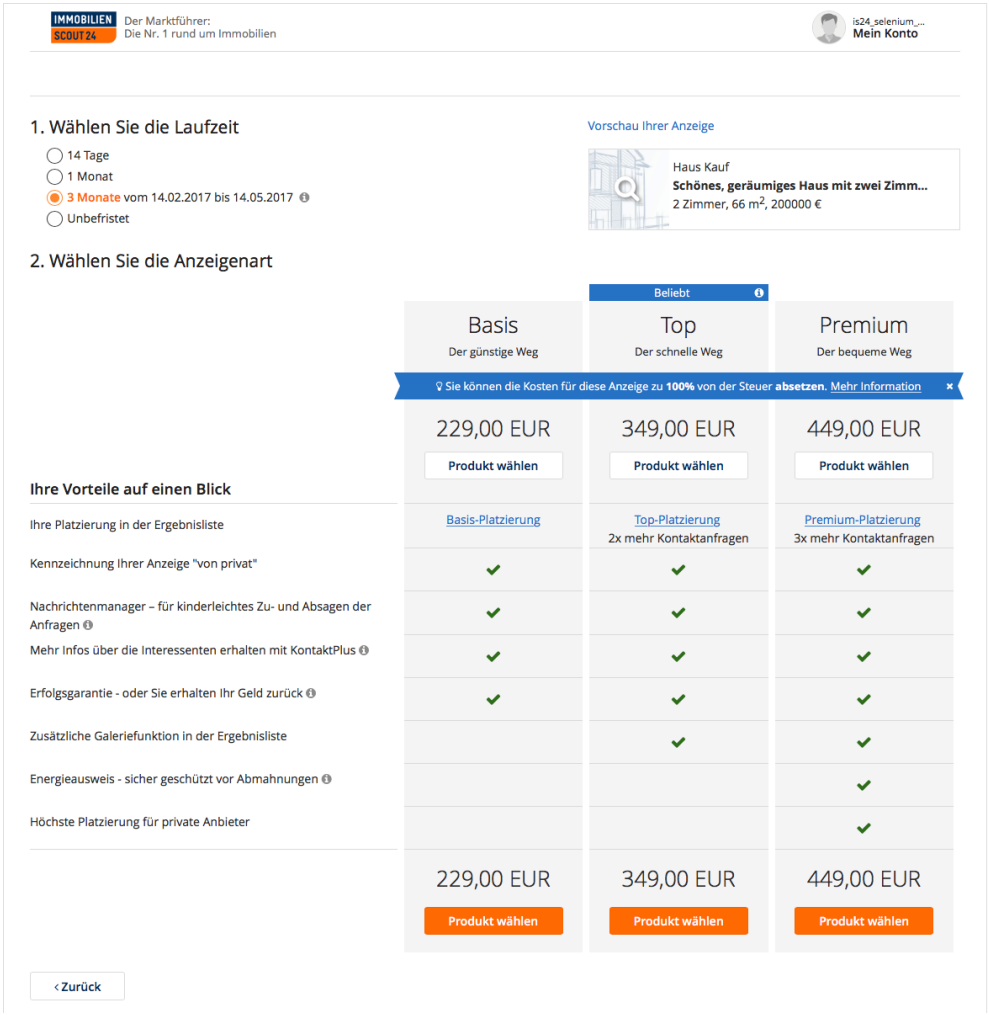

Abbildung 12: A/B-Test Steuertipp

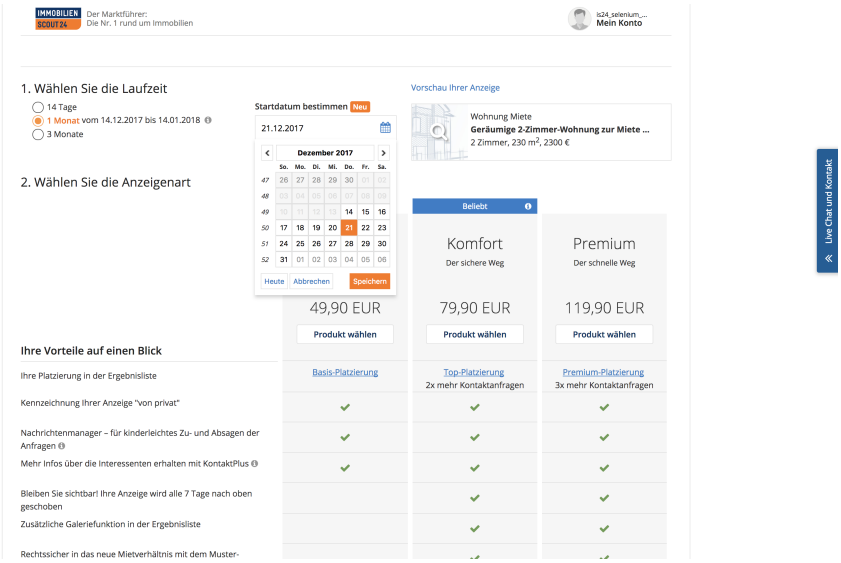

Abbildung 13: A/B-Test Startdatum wählen

rung des frei wählbaren Startzeitpunkts erhebliche Kosten verursacht hätte. Darum wurde bei der Kompromissanalyse dieses Mal ein Verhältnis von $1 : 3$ $(\alpha : \beta)$ für die Fehlerwahrscheinlichkeiten gewählt. Um den negativen Effekt auf die Kundenzufriedenheit zu reduzieren, wurde zudem erwogen, die Besucher nicht gleichmäßig auf die beiden Gruppen aufzuteilen, sondern nur jedem zehnten den Kalender-Link zu zeigen. Zu Gunsten einer kürzeren Testlaufzeit wurde dieser Gedanke allerdings verworfen.

Wegen technischer Schwierigkeiten wurde der Test später als urpsrünglich gedacht gestartet, so dass bis zum Zeitpunkt der Fertigstellung dieser Arbeit noch keine Auswertung erfolgt war.

7 Fazit

Auf den ersten Blick ist ein A/B-Test eine einfache Sache. Das Prinzip ist schnell verstanden und auch die technische Implementierung wird zunehmend unkomplizierter. Bei genauerer Betrachtung offenbaren sich jedoch etliche Fallstricke, sowohl bei der Planung als auch bei der Auswertung. Hier noch einmal die wichtigsten:

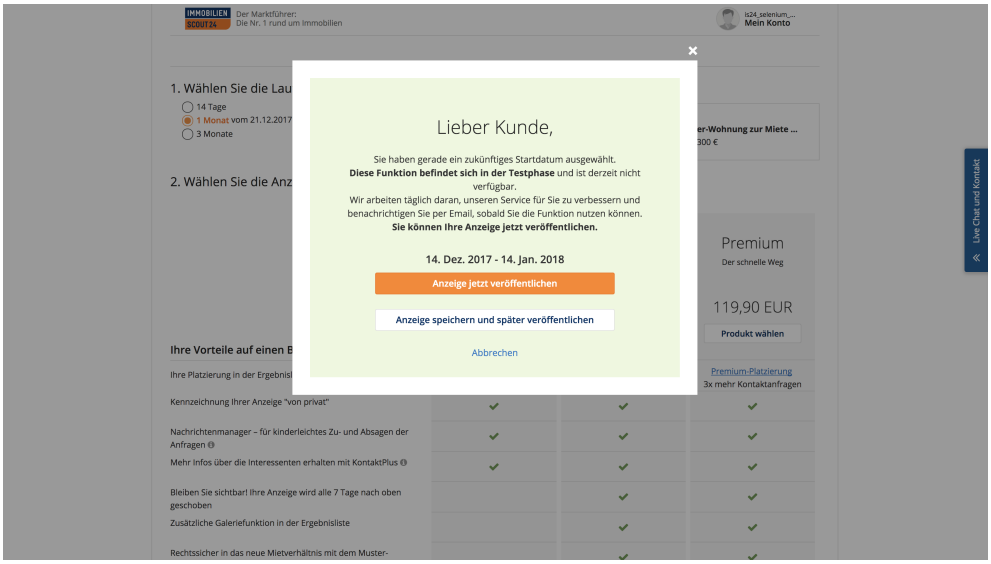

Abbildung 14: Benutzerhinweis

- **Planung:** Bei einem Signifikanzniveau von 5% kann eine akzeptable Teststärke von 80% und mehr in den meisten Fällen nur durch eine sehr große Stichprobe, mithin über eine lange Testlaufzeit erreicht werden. Dies trifft insbesondere auf Seiten zu, die bereits öfter optimiert wurden, so dass der erwartete Effekt einer neuerlichen Änderung eher gering ausfallen wird.

- **Auswertung:** Die Interpretation eines p-Werts ist weder intuitiv noch leicht eingängig. Dazu kommt: Der p-Wert ist kein Indikator der Effektstärke. Es sollte also zusätzlich ein Effektmaß berücksichtigt werden.

Diese Punkte wurden beim ABTP gezielt angegangen: Als Alternative zur klassischen Stichprobenumfangsplanung, die zu nicht realisierbaren Stichprobengrößen tendiert, wurde die Kompromissanalyse implementiert. Hier kann der Benutzer selbst festlegen, wie groß die Stichprobe sein soll. Zum Zweck der leichteren Interpretation werden die Ergebnisse dynamisch in Textform dargestellt.

Der ABTP ist mittlerweile seit ca. fünf Monaten im Einsatz. Dabei hat sich gezeigt, dass der Funktionsumfang für die meisten, aber nicht für alle Testvorhaben ausreichend ist. Wie das Beispiel in Abschnitt 6.4.2 gezeigt

hat, kann es in bestimmten Fällen sinnvoll sein, die Besucher ungleichmäßig auf die beiden Gruppen aufzuteilen. Bisher basiert der gesamte Plan-Tab jedoch auf der Annahme gleicher Gruppengrößen. Ferner ist angedacht, den ABTP noch in anderer Hinsicht zu erweitern: In Zukunft soll der ABTP auch A/B-Tests abdecken, in denen die Zielgröße ein Mittelwert ist, z. B. die durchschnittliche Anzahl der Klicks auf einen bestimmten Button pro Besucher. Dafür wird es nötig sein, einen weiteren Test einzubinden.

8 Anhang

8.1 Verteilung von Cohens w

Die Verteilung von Cohens w wurde meiner Kenntnis nach in der Literatur bisher noch nicht untersucht.[31] Wie wir in Abbildung 4 gesehen haben, liegt die Vermutung nahe, dass w asymptotisch einer Normalverteilung folgt.[32] Um dies zu prüfen, wurde eine dritte Serie von 10 000 2×2-Kontingenztafeln aus der in Tabelle 4 gezeigten Population gezogen – dieses Mal mit $n = 10\,000$. Abbildung 15 zeigt das Histogramm der beobachteten Effektstärke sowie eine Normalverteilung mit $\mu = \overline{\hat{w}}$ und $\sigma = ESD(\hat{w})$. Zum Vergleich zeigt Abbildung 16 das Histogramm einer aus dieser Verteilung gezogenen Zufallsstichprobe ($n = 10\,000$). Die große Ähnlichkeit liegt auf der Hand.

[31] Ein Ansatzpunkt findet sich bei Fleiss et al. (2003: 134), die zeigen, dass der Phi-Koeffizient, der gemäß Cohen (1988: 223) für 2×2-Tafeln dem Betrag nach w entspricht, für Werte nahe Null asymptotisch normalverteilt ist.

[32] Genau genommen muss es sich dabei um eine gefaltete Normalverteilung handeln, da per Definition $w \geq 0$ gilt.

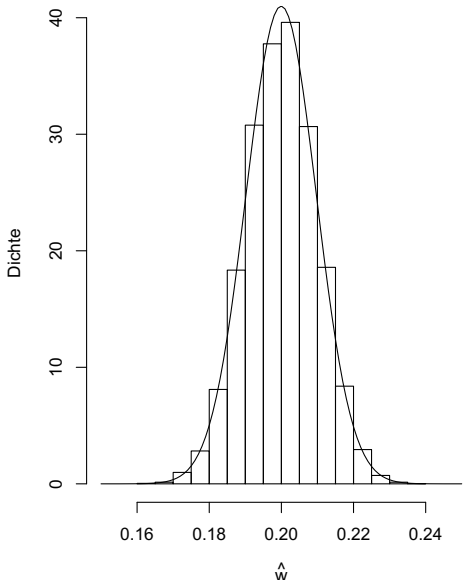

Abbildung 15: Histogramm der beobachteten Teststärke sowie Dichtefunktion einer Normalverteilung mit $\mu = \overline{\hat{w}}$ und $\sigma = ESD(\hat{w})$

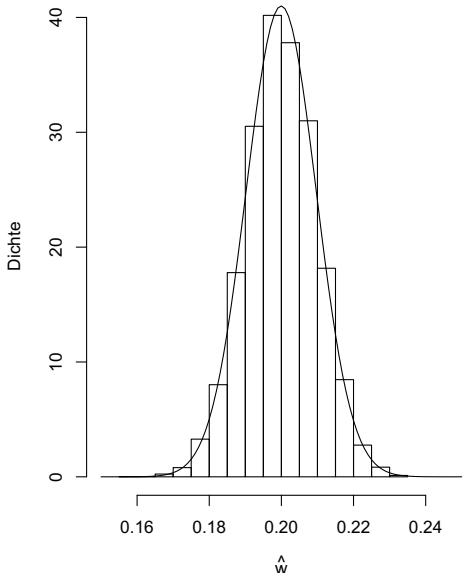

Abbildung 16: Histogramm einer Zufallsstichprobe gezogen aus einer Normalverteilung mit $\mu = \overline{\hat{w}}$ und $\sigma = ESD(\hat{w})$ sowie deren Dichtefunktion

8.2 Herleitung von Gleichung 10

Gesucht ist eine Darstellung von Gleichung 3, in der statt p_{0i} und p_{1i} die erwartete Konversionsrate der Kontrollgruppe (k) und der Effekt durch D (e) vorkommen. Bei gleich großen Gruppen gilt für die relativen Häufigkeiten unter H_1 (p_{1i}):

Tabelle 9: Relative Häufigkeiten unter H_1

Kontrollgruppe	Variante	
$0.5 \cdot k$	$0.5 \cdot k \cdot e$	$0.5 \cdot k \cdot (1 + e)$
$0.5 - 0.5 \cdot k$	$0.5 - 0.5 \cdot k \cdot e$	$1 - 0.5 \cdot k \cdot (1 + e)$
0.5	0.5	1

Folglich lauten die relativen Häufigkeiten unter H_0 (p_{0i}):

Tabelle 10: Relative Häufigkeiten unter H_0

Kontrollgruppe	Variante	
$0.25 \cdot k \cdot (1 + e)$	$0.25 \cdot k \cdot (1 + e)$	$0.5 \cdot k \cdot (1 + e)$
$0.5 - 0.25 \cdot k \cdot (1 + e)$	$0.5 - 0.25 \cdot k \cdot (1 + e)$	$1 - 0.5 \cdot k \cdot (1 + e)$
0.5	0.5	1

Setzt man diese Werte für p_{0i} und p_{1i} in Gleichung 3 ein, ergibt sich Gleichung 10.

Literatur

Banjanovic, Erin S. & Osborne, Jason W. (2016): „Confidence intervals for effect sizes. Applying bootstrap resampling“. In: *Practical Assessment, Research & Evaluation* 21, S. 1–20.

Bortz, Jürgen & Schuster, Christof (2010): *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin, Heidelberg: Springer-Verlag.

Christian, Brian (2012): „The A/B test. Inside the technology that's changing the rules of business“. URL: `https://www.wired.com/2012/04/ff_abtesting/` [Stand: 10.03.2018].

Campell, Ian (2007): „Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations“. In: *Statistics in Medicine* 26, 3661–3675.

Champely, Stephane (2016): „Pwr. Basic function for power analysis“. URL: `https://CRAN.R-project.org/package=pwr` [Stand: 10.03.2018].

Chang, Winston (2017): „Shiny. Web application framework for R“. URL: `https://CRAN.R-project.org/package=shiny` [Stand: 10.03.2018].

Chen, Li-Ting & Peng, Chao-Ying J. (2014): „The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes“. In: *Behavior Research Methods* 47, S. 107–126.

Cochran, William G. (1952): „The χ^2 test of goodness of fit“. In: *The Annals of Mathematical Statistics* 23, S. 315–345.

Cohen, Jacob (1962): „The statistical power of abnormal-social psychological research. A review“. In: *The Journal of Abnormal and Social Psychology* 65, S. 145–153.

Cohen, Jacob (1969): *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, Jacob (1988): *Statistical Power Analysis for the Behavioral Sciences*. 2. Aufl. New Jersey: Lawrence Erlbaum Associates.

Cohen, Jacob (1990): „Things I have learned (so far)“. In: *American Psychologist* 45, S. 1304–1312.

Cohen, Jacob (1992): „A power primer“. In: *Psychological Bulletin* 112, S. 155–159.

Coulter, Keith S. & Coulter, Robin A. (2005): „Size does matter. The effects of magnitude representation congruency on price Perceptions and purchase likelihood “. In: *Journal of Consumer Psychology* 15, S. 64–76.

Cumming, Geoff & Finch, Sue (2001): „A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions“. In: *Educational and Psychological Measurement* 61, S. 532–574.

Erdfelder, Edgar (1984): „Zur Bedeutung und Kontrolle des β-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle“. In: *Zeitschrift für Sozialpsychologie* 15, S. 18–32.

Erdfelder, Edgar et al. (2010): „Effektgröße und Teststärke“. In: Holling, Heinz & Schmitz, Bernd (Hg.): *Handbuch Statistik, Methoden und Evaluation*. Göttingen: Hogrefe Verlag, S. 358–369.

Falk, Ruma & Greenbaum, Charles W. (1995): „Significance tests die hard“. In: *Theory and Psychology* 5, S. 75–98.

Fleiss, Joseph L. et al. (2003): *Statistical methods for rates and proportions*. 3. Aufl. Hoboken: John Wiley & Sons.

Gigerenzer, Gerd et al. (2004): „The null ritual. What you always wanted to know about significance testing but were afraid to ask“. In: Kaplan, David (Hg.): *The SAGE handbook of quantitative methodology for the social sciences*. Thousand Oaks u. a.: SAGE Puclications, S. 391–408.

Goodman, Steven (2008): „A dirty dozen. Twelve p-value misconceptions“. In: *Seminars in Hematology* 45, S. 135–140.

Hassler, Marco (2012): *Web Analytics. Metriken auswerten, Besucherverhalten verstehen, Website optimieren*. 3. Aufl. Heidelberg u. a.: Verlagsgruppe Hüthig Jehle Rehm.

Hern, Alex (2014): „Why Google has 200m reasons to put engineers over designers“. URL: `https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers` [Stand: 10.03.2018].

Hoenig, John M. & Heisey, Dennis M. (2001): „The abuse of power. The pervasive fallacy of power calculations for data analysis“. In: *The American Statistician* 55, S. 19–24.

Ioannidis, John P. A. (2005): „Why most published research findings are false“. In: *PLOS Medicine* 2, S. 696–701.

Ioannidis, John P. A. (2008): „Why most discovered true associations are inflated“. In: *Epidemiology* 19, S. 640–648.

Kelley, Ken (2005): „The effects of nonnormal distributions on confidence intervals around the standardized mean difference. Bootstrapping as an alternative to parametric confidence intervals“. In: *Educational and Psychological Measurement* 65, S. 51–69.

Kohvai, Ron et. al. (2008): „Controlled experiments on the web. Survey and practical guide“. In: *Data Mining and Knowledge Discovery* 18, S. 140–181.

Krüger, Jörg D. (2011): *Conversion Boosting mit Website Testing*. Heidelberg u. a.: Verlagsgruppe Hüthig Jehle Rehm.

Monetate (2017): *Monetate Ecommerce Quarterly Report for Q2 2017*. URL: `http://info.monetate.com/ecommerce-quarterly-report-q2-2017.html` [Stand: 10.03.2018].

R Core Team (2016): *R. A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

R Foundation for Statisctical Computing (o. D.): *Contributed Packages*. URL: `https://cran.r-project.org/` [Stand: 10.03.2018].

Ryan, Thomas P. (2013): *Sample size determination and power*. Hoboken: John Wiley & Sons.

Sedlmeier, Peter & Gigerenzer, Gerd (1989): „Do studies of statistical power have an effect on the power of studies?“. In: *Psychological Bulletin* 105, S. 309–316.

Soetaert, Karline (2009): „RootSolve. Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations“. URL: `https://CRAN.R-project.org/package=rootSolve` [Stand: 10.03.2018].

Steiger, James H. & Fouladi, Rachel T. (1997): „Noncentrality interval estimation and the evaluation of statistical models“. In: Harlow, Lisa L. et al. (Hg.): *What if there were no significance tests?*. New Jersey: Lawrence Erlbaum Associates, S. 221–257.

Sterne, Jonathan A C & Smith, George D. (2002): „Sifting the evidence. What's wrong with significance tests?“. In: *BMJ* 322, S. 226–231.

Stigler, Stephen (2008): „Fisher and the 5% level“. In: *Chance* 21, S. 12.

Thomas, Len (1997): „Retrospective power analysis“. In: *Conservation Biology* 11, S. 276–280.

Wickham, Hadley (2014): *Advanced R*. Boca Raton: Taylor & Francis Group.

Yuan, Ke-Hai & Maxwell, Scott (2005): „On the post hoc power in testing mean differences“. In: *Journal of Educational and Behavioral Statistics* 30, S. 141–167.

Ehrenwörtliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst habe. Gedanken anderer sind als solche kenntlich gemacht. Die Arbeit wurde in gleicher oder änlicher Form bisher keiner anderen Prüfungsbehörde vorgelegt.

Berlin, 10. März 2018